

VERİ MADENCİLİĞİ

Giriş

Prof. Dr. Şule Gündüz Öğüdücü

1

Konular

- Veri madenciliği ve bilgi keşfinin tanımı
- Veri madenciliğinin uygulama alanları
- Veri kaynakları
- Veri madenciliği modellerinin gruplanması
- Veri ambarları
- Veri madenciliğinde sorunlar

2

Problem Tanımı

- teknolojinin gelişimiyle bilgisayar ortamında ve veri tabanlarında tutulan veri miktarının artması (terabyte -> petabyte)
 - verinin kolayca toplanabilmesi
 - bu veriyi nasıl kullanacağımızı bilmiyoruz
 - saklanan veriden bilgi elde etmek için bu veriyi yorumlamamız gerekiyor
- kullanıcıların beklentilerinin artması
 - basit veri tabanı sorgulama yöntemlerinin yeterli olmaması
- Veri madenciliği yöntemleri fazla miktardaki veri içinden yararlı bilgiyi bulmak için kullanılır.



3

Veri Madenciliği: Tarihçe

- Data Fishing-Data Dredging: 1960-
 - istatistikçiler
- Data Mining: 1990-
 - veritabanı kullanıcıları, ticari
- Knowledge Discovery in Databases (KDD): 1989-
 - yapay zeka, makine öğrenmesi toplulukları
- Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction,...



4

Bilgi Keşfi

- Büyük veri kaynaklarından yararlı ve ilginç bilgiyi bulmak
- Bulunan bilgi
 - gizli
 - önemli
 - önceden bilinmeyen
 - yararlı



5

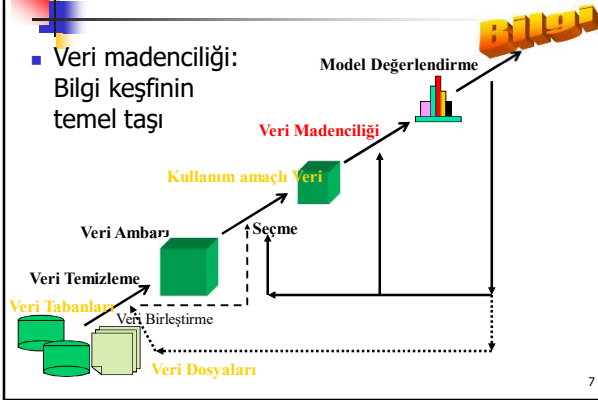
Veri Madenciliği Nedir?

- Teoride veri madenciliği bilgi keşfi işleminin aşamalarından biridir.
- Pratikte veri madenciliği ve bilgi keşfi eş anlamlı olarak kullanılır.
- Veri madenciliği teknikleri veriyi belli bir modele uydurur.
 - veri içindeki *örüntüleri* bulur
 - örüntü: veri içindeki herhangi bir yapı
- Sorgulama ya da basit istatistik yöntemler veri madenciliği değildir.

6

Bilgi Keşfi

- Veri madenciliği: Bilgi keşfinin temel taşı



7

Bilgi Keşfinin Aşamaları

- Uygulama alanını inceleme
 - Konuyla ilgili bilgi ve uygulama amaçları
- Amaca uygun veri kümesi oluşturma
- Veri ayıklama ve ön işleme
- 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

işlemin %70'lik bölümünü oluşturur



8

Bilgi Keşfi Örnek: web kayıtları

- web sitesinin yapısını inceleme
- verileri seçme: tarih aralığını belirleme
- veri ayıklama, ön işleme: gereksiz kayıtları silme
- veri azaltma, veri dönüşümü: kullanıcı oturumları belirleme
- veri madenciliği tekniği seçme: demetleme
- veri madenciliği algoritmasını seçme: k-ortalama, EM, DBSCAN...
- Model değerlendirme/yorumlama: değişik kullanıcı grupları için sıkça izlenen yolu bulma
- Uygulama alanları: öneri modelleri, kişiselleştirme, ön belleğe alma

9

Bilgi Keşfi: Farklı Alanların Bileşimi



10

Veri Madenciliğine Farklı Bakış Açıları

- Veri madenciliği uygulanacak veri
 - veri dosyaları, ilişkisel veritabanları, veri ambarları, nesneye dayalı veritabanları, www, konumsal veriler, zamansal veriler
- Veri madenciliği işlevleri
 - demetleme, sınıflandırma, ilişkilendirme kuralları, eğri uydurma, kestirim, aykırılıkların belirlenmesi..
 - veri içindeki örüntülerin bulunması /aykırılıkların bulunması
- Kullanılan teknikler
 - kullanıcıyla etkileşim seviyesi
 - veri analizi için kullanılan yöntemler (makine öğrenmesi, istatistik, görüntüleme...)
- Uygulama alanı
 - telekomunikasyon, finans, sahtekarlık belirlenmesi, biyolojik veri madenciliği, web madenciliği

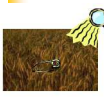
11

İstatistik & Makine Öğrenmesi & Veri Madenciliği

- İstatistik
 - daha çok teoriye dayalı yaklaşımlar
 - bir varsayımın doğruluğunu araştırır
- Makine Öğrenmesi
 - daha çok sezgisel yaklaşımlar
 - öğrenme işleminin başarımını artırmaya çalışır
- Veri madenciliği ve bilgi keşfi
 - teori ve sezgisel yaklaşımları birleştirir
 - bilgi keşfinin tüm aşamalarını gerçekleştirir: veri temizleme, öğrenme, sonucu sunma, yorumlama,...
- Aradaki ayrım net değil

12

Veri Madenciliğinin Amacı



- Yapmak **istemediğimiz**: Büyük miktardaki veri içinde arama yapmak (Veri tabanı yönetim sistemleri bu işi yapıyor)



- Amaç: Aradığımız veri mevcutsa sonuçlarını anlamak

13

Veritabanı & Veri Madenciliği İşlemleri

■ Sorgulama

- Tanımlı
- SQL

■ Veri

- Canlı veri

■ Çıkış

- Belirli
- verinin bir alt kümesi

■ Sorgulama

- Tam tanımlı değil
- yaygın sorgulama dili yok

■ Veri

- Üzerinde işlem yapılmayan veri

■ Çıkış

- Belirli değil
- verinin bir alt kümesi değil

14

Sorgulama Örnekleri

■ Veritabanı uygulaması:

- Soyadı Gündüz olan kredi kartı sahiplerini bul.
- Bir ayda 2000 YTL'den fazla harcama yapan kredi kartı sahiplerini bul.
- Önce bilgisayar daha sonra yazıcı satın alan tüm müşterileri bul.

■ Veri madenciliği uygulaması

- Riski az olan tüm kredi kartı başvurularının özelliklerini bul (sınıflandırma)
- Harcama alışkanlığı benzer olan kredi kartı sahiplerinin özelliklerini bul (demetleme)
- Sıkça satın alınan ürünleri bul (ilişkilendirme kuralları)

15

Konular

- Veri madenciliği ve bilgi keşfinin tanımı
- Veri madenciliğinin uygulama alanları
- Veri kaynakları
- Veri madenciliği modellerinin gruplanması
- Veri ambarları
- Veri madenciliğinde sorunlar

16

Veri Madenciliğinin Uygulama Alanları

■ Veritabanı analizi ve karar verme desteği

- Pazar araştırması
 - Hedef Pazar, müşteriler arası benzerliklerin saptanması, sepet analizi, çapraz pazar incelemesi
- Risk analizi
 - Rekabet analizi, öngörü
- Sahtekarlıkların saptanması

■ Diğer Uygulamalar

- Belgeler arası benzerlik (haber kümeleri, e-posta)
- Sorgulama sonuçları

17

Uygulamalar

- Hangi promosyonu ne zaman uygulamalıyım?
- Hangi müşteri aldığı krediyi geri ödemeyebilir?
- Bir müşteriye ne kadar kredi verilebilir?
- Sahtekarlık olabilecek davranışlar hangileridir?
- Hangi müşteriler yakın zamanda kaybedilebilir?
- Hangi müşterilere promosyon yapmalıyım?
- Hangi yatırım araçlarına yatırım yapmalıyım?

18

Pazar Araştırması (1)

- Veri madenciliği uygulamaları için veri kaynağı
 - Kredi kartı hareketleri, üyelik kartları, ucuzluk kuponları, pazar anketleri
- Hedef pazarlar bulma
 - Benzer özellikler gösteren müşterilerin bulunması: benzer gelir grupları, ilgi alanları, harcama alışkanlıkları
- Müşterilerin davranışlarında zaman içindeki değişiklik
 - Tek kişilik banka hesabının ortak hesaba çevrilmesi: evlilik
- Çapraz pazar incelemesi:
 - Ürün satışları arasındaki ilişkiyi bulma

19

Pazar Araştırması (2)

- Müşteri profili
 - Hangi özellikteki müşterilerin hangi ürünleri aldıkları (demetleme veya sınıflandırma)
- Müşterilerin ihtiyaçlarını belirleme
 - Farklı müşterilerin o anki ilgisine yönelik ürünü bulma
 - Yeni müşterileri hangi faktörlerin etkilediğini bulma

20

Sahtekarlık İncelemesi

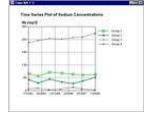
- Sigorta, bankacılık, telekomünikasyon alanlarında
- Geçmiş veri kullanılarak sahtekarlık yapanlar için bir model oluşturma ve benzer davranış gösterenleri belirleme
- Örnek:
 - Araba sigortası
 - Sağlık Sigortası
 - Kredi kartı başvurusu



21

Risk Analizi

- Finans planlaması ve bilanço değerlendirmesi
 - nakit para akışı incelemesi ve kestirimi
 - talep incelemesi
 - zaman serileri incelemesi
- Kaynak planlaması
 - kaynakların incelenmesi ve uygun olarak dağıtılması
- Rekabet
 - rakipleri ve pazar eğilimlerini takip etme
 - müşterileri sınıflara ayırma ve fiyat politikası belirleme



22

Konular

- Veri madenciliği ve bilgi keşfinin tanımı
- Veri madenciliğinin uygulama alanları
- Veri kaynakları
- Veri madenciliği modellerinin gruplanması
- Veri ambarları
- Veri madenciliğinde sorunlar

23

Veri Kaynakları

- Veri dosyaları
- Veritabanı kaynaklı veri kümeleri
 - ilişkisel veritabanları, veri ambarları
- Gelişmiş veri kümeleri
 - duraksız veri (data stream), algılayıcı verileri (sensor data)
 - zaman serileri, sıralı diziler (biyolojik veriler)
 - çizgeler, sosyal ağ (social networks) verileri
 - konumsal veriler (spatial data)
 - çokul ortam veritabanları (multimedia databases)
 - nesneye dayalı veritabanları
 - www



24

Konular

- Veri madenciliği ve bilgi keşfinin tanımı
- Veri madenciliğinin uygulama alanları
- Veri kaynakları
- Veri madenciliği modellerinin gruplanması
- Veri ambarları
- Veri madenciliğinde sorunlar

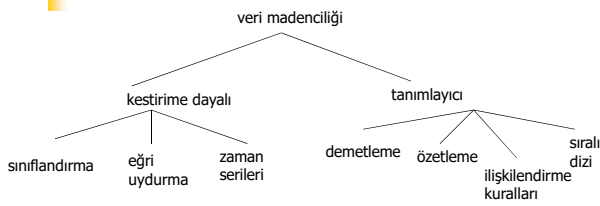
25

Veri Madenciliği Algoritmaları

- amaç: veriyi belli bir modele uydurmak
 - tanımlayıcı
 - En iyi müşterilerim kimler?
 - Hangi ürünler birlikte satılıyor?
 - Hangi müşteri gruplarının alışveriş alışkanlıkları benzer?
 - kestirime dayalı
 - Kredi başvurularını risk gruplarına ayırma
 - Şirketle çalışmayı bırakacak müşterileri öngörme
 - Borsa tahmini
- seçim: veriye uyan en iyi modeli seçmek için kullanılan kriter
- arama: veri üzerinde arama yapmak için kullanılan teknik

26

Veri Madenciliği Modelleri



27

Veri Madenciliği İşlevleri

- Sınıflandırma (Classification): Veriyi önceden belirlenmiş sınıflardan birine dahil eder.
 - Gözetimli öğrenme
 - Örüntü tanıma
 - Kestirim
- Eğri uydurma (Regression): Veriyi gerçek değerli bir fonksiyona dönüştürür.
- Zaman serileri inceleme (Time Series Analysis): Zaman içinde değişen verinin değerini öngörür.
- Aykırlıkları Belirleme (Outlier Analysis): Verinin geneline uymayan nesneleri belirleme

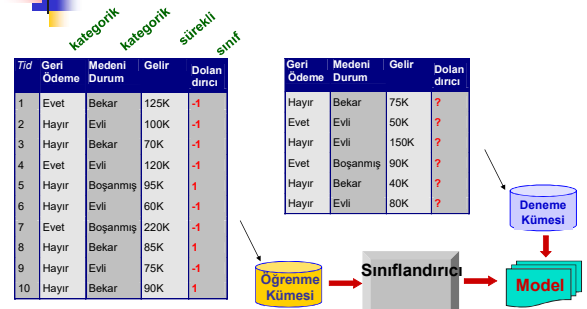
28

Veri Madenciliği İşlevleri

- Demetleme (Clustering): Benzer verileri aynı grupta toplama
 - Gözetimsiz öğrenme
- Özetleme (Summarization): Veriyi alt gruplara ayırır. Her alt grubu temsil edecek özellikler bulur.
 - Genelleştirme (Generalization)
 - Nitelendirme (Characterization)
- İlişkilendirme kuralları (Association Rules)
 - Veriler arasındaki ilişkiyi belirler
- Sıralı dizileri bulma (Sequence Discovery): Veri içinde sıralı örüntüler bulmak için kullanılır.

29

Örnek: Sınıflandırma



30

Örnek: Demetleme

- Doküman Demetleme:
- Amaç:
 - Doküman içinde geçen terimlere göre aynı konudaki dokümanları gruplama
- Yaklaşım:
 - Her doküman içinde sık geçen terimleri bul. Bu terimlerden ve ağırlıklarından yararlanarak bir benzerlik ölçütü geliştir. Bu ölçüte göre demetleme yap
- Kullanımı:
 - Yeni bir dokümanın hangi dokümanlarla benzer olduğu
 - terimlere göre arama yapıldığında bu terimleri içeren dokümanları bulma

31

Örnek: İlişkilendirme Kuralları

- Veri kümesindeki nesneler arasındaki ilişkiyi bulma
 - bir nesnenin (nesnelerin) varlığı ile diğer bir nesnenin (nesnelerin) de varlığını tahmin edebilecek kurallar

TID	Nesneler
1	Ekmek, Kola, Süt
2	Bira, Ekmek
3	Bira, Kola, Çocuk bezi, Süt
4	Bira, Ekmek, Çocuk bezi, Süt
5	Kola, Çocuk bezi, Süt

Bulunan Kurallar:
 {Süt} --> {Kola}
 {Çocuk bezi, Süt} --> {Bira}

32

Bulunan Örüntüler Önemli mi?

- Binlerce örüntü: Bir kısmı önemli
 - Veri madenciliği ile bulunan sonuç kümesi üzerinde tekrar veri madenciliği uygulanacak kadar büyük
- Bulunan örüntünün önemli olması için:
 - insanlar tarafından *kolayca anlaşılabilir*
 - sına verisi veya yeni veriler üzerinde belli oranda *geçerli*
 - *yararlı* ve kullanılabilir
 - *yeni*
- nesnel / öznel metrikler
 - nesnel: örüntünün yapısına bağlı
 - öznel: kullanıcının yaklaşımına bağlı

33

Bütün Önemli Örüntülerin Bulunması

- Bütünlük: Önemli bütün örüntülerin bulunması
 - Önemli bütün örüntüler bulunabilir mi?
- Eniyileme: Sadece önemli örüntüleri bulma
 - Sadece önemli örüntüler bulunabilir mi?
 - Bütün örüntüler bulunduktan sonra önemsiz olanların ayıklanması
 - Sadece önemli örüntülerin bulunması

34

En İyi 10 Veri Madenciliği Algoritması Belirlenen 18 Aday

- Classification
 - #1. C4.5 Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc.
 - #2. CART L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
 - #3. K Nearest Neighbours (kNN) Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616.
 - #4. Naive Bayes Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- Statistical Learning
 - #5. SVM Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.
 - #6. EM McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York.
- Association Analysis
 - #7. Apriori: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
 - #8. FP-Tree: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.

35

En İyi 10 Veri Madenciliği Algoritması Belirlenen 18 Aday

- Link Mining
 - #9. PageRank: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
 - #10. HITS: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. SODA, 1998.
- Clustering
 - #11. K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
 - #12. BIRCH: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
- Bagging and Boosting
 - #13. AdaBoost: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.

36

En İyi 10 Veri Madenciliği Algoritması Belirlenen 18 Aday

- Sequential Patterns
 - #14. GSP: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
 - #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- Integrated Mining
 - #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- Rough Sets
 - #17. Finding reduct: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- Graph Mining
 - #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

37

Seçilen En İyi 10 Veri Madenciliği Algoritması ICDM'06

- #1: C4.5 (61 votes)
- #2: K-Means (60 votes)
- #3: SVM (58 votes)
- #4: Apriori (52 votes)
- #5: EM (48 votes)
- #6: PageRank (46 votes)
- #7: AdaBoost (45 votes)
- #7: kNN (45 votes)
- #7: Naive Bayes (45 votes)
- #10: CART (34 votes)

38

Bu Konudaki Başlıca Konferanslar ve Dergiler

- KDD konferansları
 - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
 - SIAM Data Mining Conf. (SDM)
 - (IEEE) Int. Conf. on Data Mining (ICDM)
 - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
 - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Konuyla ilgili diğer konferanslar
 - ACM SIGMOD
 - VLDB
 - (IEEE) ICDE
 - WWW, SIGIR
 - ICML, CVPR, NIPS
- Dergiler
 - Data Mining and Knowledge Discovery (DAMI or DMKD)
 - IEEE Trans. On Knowledge and Data Eng. (TKDE)
 - KDD Explorations
 - ACM Trans. on KDD

39

Konular

- Veri madenciliği ve bilgi keşfinin tanımı
- Veri madenciliğinin uygulama alanları
- Veri kaynakları
- Veri madenciliği modellerinin gruplanması
- Veri ambarları
- Veri madenciliğinde sorunlar

40

İlgili Konular: Veri Ambarları

- Çok fazla miktarda üzerinde işlem yapılan veri var
- Çoğunlukla farklı veri tabanlarında ve farklı ortamlarda
- Veri farklı formatlarda ve yerlerde (heterojen ve dağınık)



- Karar destek birimleri veriye sanal olarak tek bir yerden ulaşabilmeli
- Ulaşım hızlı olmalı

41

Veri Ambarı

- Amaca yönelik
- Birleştirilmiş
- Zaman değişkenli
- Değişken değil

42

Veri Ambarları: Amaca Yönelik

- Müşteri, ürün, satış gibi belli konular için düzenlenebilir
- Verinin incelenmesi ve modellenmesi için oluşturulur
- Konuyla ilgili karar vermek için gerekli olmayan veriyi kullanmayarak konuya basit, özet bakış sağlar

43

Veri ambarları: Birleştirilmiş

- Veri kaynaklarının birleştirilmesiyle oluşturulur
 - Canlı veri tabanları, dosyalar
- Veri temizleme ve birleştirme teknikleri kullanılır
 - Değişik veri kaynakları arasındaki tutarlılık sağlanır

44

Veri Ambarları: Zaman Değişkenli

- Zaman değişkeni canlı veri tabanlarına göre daha uzundur
 - Canlı veri tabanları: Güncel veriler bulunur (en çok geçmiş 1 yıl)
 - Veri ambarları: Geçmiş hakkında bilgi verir (geçmiş 5-10 yıl)

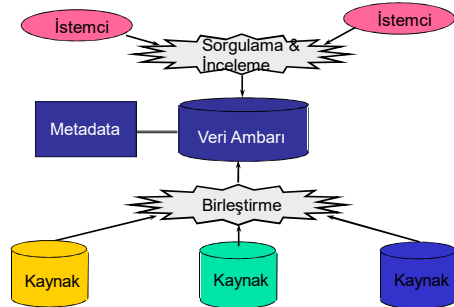
45

Veri Ambarları: Değişken Değil

- Canlı veri tabanlarından alınmış verinin fiziksel olarak başka bir ortamda saklanması
- Canlı veri tabanlarındaki değişimin veri ambarlarını etkilememesi

46

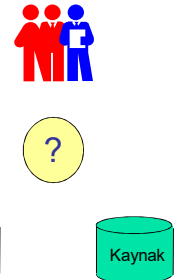
Veri Ambarı Mimarisi



47

Veri Kaynakları

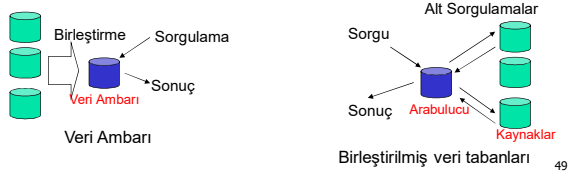
- İki yaklaşım:
 - sorgulamalı
 - veri ambarı



48

Veri Ambarı & Birleşmiş Veritabanları

- Veri tabanlarının birleştirilmesi:
 - Farklı veri tabanları arasında bir arabulucu katman
 - Sorgulamalı
 - Bir sorgulamayı her veri tabanı için alt sorgulamalara ayır
 - Sonucu birleştir
- Veri ambarı:
 - Veri daha sonra kullanılmak üzere birleştirilip veri ambarında saklanıyor



49

Veri Madenciliği & OLAP

- OLAP (On-Line Analytical Processing)
 - Veri ambarlarının işlevi
 - Veriyi inceleme ve karar verme
- OLTP (On-Line Transaction Processing) saatler sürebilen işlemler
- OLAP avantajları
 - Daha geniş kapsamlı sonuçlar
 - Daha kısa süreli işlem
- OLAP dezavantajları
 - Kullanıcı neyi nasıl soracağını bilmesi gerekiyor
 - Genelde veriden istatistiksel inceleme yapmak için kullanılır.

OLAP NE sorusuna cevap verir, veri madenciliği NEDEN sorusuna cevap verir.

50

Konular

- Veri madenciliği ve bilgi keşfinin tanımı
- Veri madenciliğinin uygulama alanları
- Veri kaynakları
- Veri madenciliği modellerinin gruplanması
- Veri ambarları
- Veri madenciliğinde sorunlar

51

Veri Madenciliğinde Sorunlar (1)

- Gizlilik ve sosyal haklar
- Kullanıcı Arabirimi
- Veri madenciliği yöntemi
- Başarım ve ölçeklenebilirlik
- Veri kaynağı

52

Veri Madenciliğinde Sorunlar (2)

- Gizlilik ve sosyal haklar
 - Kişilere ait verilerin toplanarak, kişilerden habersiz ve izinsiz olarak kullanılması
- Veri madenciliği yöntemleri ile bulunan sonuçların izinsiz olarak açıklanması (/paylaşılması)
- Gizlilik ve veri madenciliği politikalarının düzenlenmesi

53

Veri Madenciliğinde Sorunlar (3)

- Kullanıcı Arabirimi
 - Görüntüleme
 - Sonucun anlaşılabilir ve yorumlanabilir hale getirilmesi
 - Bilginin sunulması
 - Etkileşim
 - Veri madenciliği ile elde edilen bilginin kullanılması
 - Veri madenciliği yöntemine müdahale etmek
 - Veri madenciliği yönteminin sonucuna müdahale etmek

54

Veri Madenciliğinde Sorunlar (4)

- Veri madenciliği yöntemi
 - Farklı tipte veriler üzerinde çalışabilme
 - Farklı seviyelerde kullanıcı ile etkileşim halinde olabilme
 - Uygulama ortamı bilgisini kullanabilme
 - Veri madenciliği ile elde edilen sonucu anlaşılır şekilde sunabilme
 - Gürültülü ve eksik veri ile çalışabilme (ve iyi sonuç verebilme)
 - Değişen veya eklenen verileri kolayca kullanabilme
 - Örüntü değerlendirme: önemli örüntüleri bulma

55

Veri Madenciliğinde Sorunlar (5)

- Başarım ve ölçeklenebilirlik
 - Kullanabilirlik ve ölçeklenebilirlik
 - Zaman karmaşıklığı ve yer karmaşıklığı kabul edilebilir
 - Örnekleme yapabilme
 - Paralel ve dağıtık yöntemler
 - Artımlı veri madenciliği
 - Parçala ve çöz

56

Uyarı

- Veri madenciliği yöntemleri bilinçsiz olarak kullanılmamalı
 - Veri madenciliği yöntemleri geçmiş olaylara bakarak örüntüler bulur: Gelecekteki olaylar geçmiştekilerle aynı değildir
 - İlişkiler her zaman nedenleri açıklamaz
 - David Heckerman'ın verdiği örnek (1997)
 - hamburger, hot-dog, barbecue sauce
 - %33 hamburger, %33 hot-dog, %33 hamburger+hot-dog
 - sadece hamburger alanlar barbecue sauce alıyor
 - Hot-dog -> barbecue sauce

57