Mineração de Dados

Introdução

1

Mineração de Dados

- Introdução
- Definições
- Descoberta de Conhecimento em Base de Dados
- Aplicações

2

Algumas Perguntas

- Que livros um cliente da Amazon gostaria de comprar ?
- É seguro dar crédito em dinheiro a uma determinada pessoa ?
- É possível detectar o roubo de um cartão de crédito pelo seu uso ?

Mas ...

O que são dados ?

O Modelo DIKW

Where is the Life we have lost in living ?
Where is the wisdom we have lost in knowledge ?
Where is the knowledge we have lost in information ?

T. S. Elliot, "The Rock", Faber & Faber, 1934.

Definição

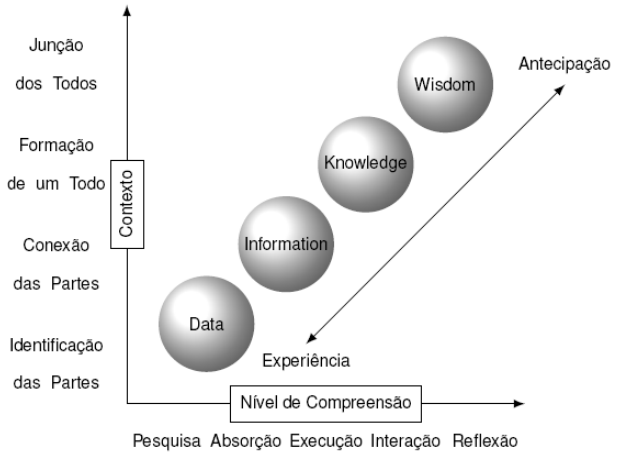
- DIKW
  - Data (Dados)
  - Information (Informação)
  - Knowledge (Conhecimento)
  - Wisdom (Sabedoria)
- Hierarquia relacionada a conceitos sobre conhecimento
- Conceitos são encadeados definindo
  - Contexto
  - Compreensão

Definição

- DIKW
  - Os seus componentes em ordem crescente de importância:
    - **Dados** (*Data*) é o nível mais básico;
    - **Informação** (*Information*) acrescenta contexto e significado aos dados;
    - **Conhecimento** (*Knowledge*) acrescenta a forma como usar adequadamente a informação;
    - **Sabedoria** (*Wisdow*) acrescenta o entendimento de quando utilizá-los.
- Desta forma, a hierarquia DIKW é um modelo teórico que se mostra útil na análise e no entendimento da importância e limites das atividades dos ***trabalhadores do conhecimento***.

7

Representação por Eixo



8

Dados e Informação

- Dados: um símbolo, um fato ou evento sem relação com outras coisas.
  - Ex. Neva
- Informação: compreensão de uma relação ou contextualização de dados. Por exemplo, uma relação de causa e efeito.
  - Ex. A temperatura caiu 8 graus e então começou a nevar.

9

Conhecimento e Sabedoria

- Compreensão de um padrão que normalmente permite deduzir o que acontecerá ou identificar algo descrito por um conjunto de fatos ou símbolos.
  - Ex. Se a umidade do ar está muito elevada e a temperatura cai a 40C, então a atmosfera muito provavelmente não consegue segurar as gotículas de água e então estas gotículas congelam durante a queda e neva.
- Agrega mais que um entendimento de princípios fundamentais encorporados ao conhecimento que são essencialmente a base para o conhecimento que faz o que ele é.

10

A origem da “Data Information Knowledge Wisdom”

- Frank Zappa fez alusão a hierarquia em 1979 [*"Packard Goose" in album Joe's Garage: Act II & III* (Tower Records, 1979)]:
  - *Information is not knowledge,*
  - *Knowledge is not wisdom,*
  - *Wisdom is not truth,*
  - *Truth is not beauty,*
  - *Beauty is not love,*
  - *Love is not music,*
  - *and Music is THE BEST.*

11

Hierarquia de Ackoff: Data-Information-Knowledge-Understanding & Wisdom

- Compreensão exige diagnótico e prescrição, que considera serem mais que "conhecimento" menos que sabedoria.
- Enquanto as informações agem rapidamente, o conhecimento tem uma vida mais longa e compreensão tem apenas uma aura de permanência.
- Sabedoria é considerada como "permanente" no verdadeiro sentido.

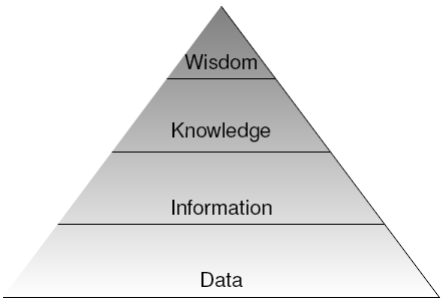
12

DIKW

December 1982 issue of THE FUTURIST.

13

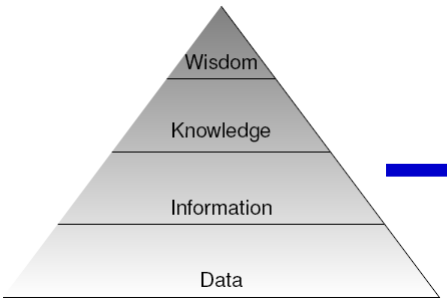
Pirâmide de Conhecimento



Renda(R), Despesas(D)

14

Pirâmide de Conhecimento

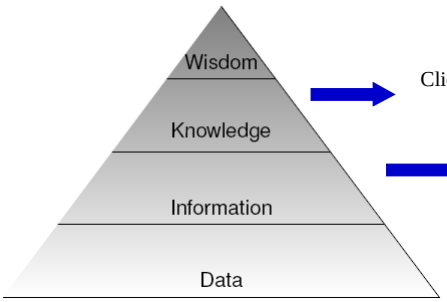


$$Capac\_Endiv = \frac{100(R-D)}{R}$$

Renda(R), Despesas(D)

15

Pirâmide de Conhecimento



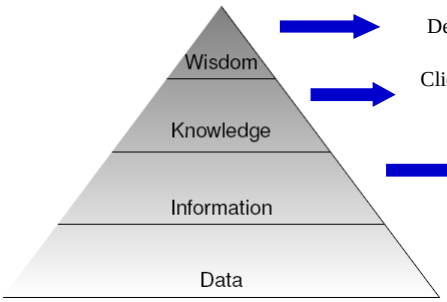
Clientes com capacidade de endividamento maior que 60% têm crédito imediato

$$Capac\_Endiv = \frac{100(R-D)}{R}$$

Renda(R), Despesas(D)

16

Pirâmide de Conhecimento



Decisão de dar ou não crédito ao cliente

Clientes com capacidade de endividamento maior que 60% têm crédito imediato

$$Capac\_Endiv = \frac{100(R-D)}{R}$$

Renda(R), Despesas(D)

17

Considere:

- I have a box.
- The box is 3' wide, 3' deep, and 6' high.
- The box is very heavy.
- The box has a door on the front of it.
- When I open the box it has food in it.
- It is colder inside the box than it is outside.
- You usually find the box in the kitchen.
- There is a smaller compartment inside the box with ice in it.
- When you open the door the light comes on.
- When you move this box you usually find lots of dirt underneath it.
- Junk has a real habit of collecting on top of this box.
- What is it?

18

Pirâmide de Conhecimento

- Os dados são os elementos através dos quais é possível se obter informação
- Com informação é possível construir **conhecimento**

Mineração de Dados
Através da investigação dos dados podemos chegar a descoberta do conhecimento

Mineração de Dados

- Não há consenso sobre definição
  - Termo recente aplicado a confluência de idéias de estatística e ciência da computação
  - Terminologia também não é padronizada
- Definições podem ser restritas ou abrangentes
  - Estatística em grandes bases de dados
  - Reconhecimento de padrões
  - Descoberta de conhecimento

Definição do Termo

"Mineração de Dados é o processo de descoberta de novas e significativas correlações, padrões e tendências em grandes volumes de dados, através do uso de técnicas e reconhecimento de padrões, estatística e outras ferramentas matemáticas."

Gartner Group

Definição do Termo

- Multidisciplinaridade
  - Estatística
  - Aprendizado de Máquina e Inteligência Computacional
  - Banco de Dados
  - Reconhecimento de Padrões

Outra Definição

- Descoberta de novos padrões em bancos de dados
  - Padrões devem ser úteis (novos e válidos)
  - Padrões podem ser inesperados
- Pode envolver um conjunto de técnicas auxiliares
  - Limpeza de dados
  - Visualização
  - Warehousing
- Podemos agregar essas técnicas auxiliares em uma definição mais abrangente ?

Podemos Estender a Definição ?

- A descoberta do conhecimento se resume apenas a análise dos dados ?
  - No mundo real os dados não estão prontos para serem prontamente analisados
  - Ser humano necessita de formas intuitivas para visualizar resultados

Data Warehouse

Data Mining

Na grande empresa, a memória é o *data warehouse*, enquanto a inteligência é o *data mining*

25

Descoberta de Conhecimento em BD

KDD: Knowledge Discovery in Database

Existem nomes tais como: *knowledge discovery in database, data mining, knowledge extraction, information discovery, data archaelogy, information harvesting e ainda data pattern processing*

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

U. M. Fayyad et al., "The KDD Process for Extracting Useful Knowledge from Volumes of Data", 1996

26

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

27

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

28

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

29

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

30

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

Fatores Operacionais

- Volumes de dados grandes e heterogêneos
- Tratamento de resultados em diferentes formatos
- Integração de algoritmos específicos

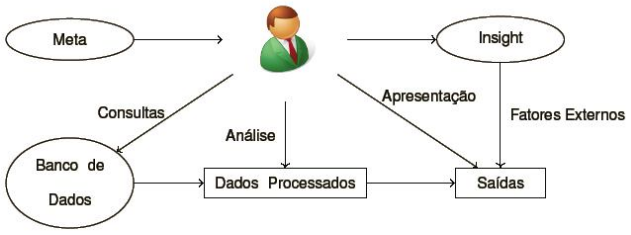
Fatores de Controle

- Formulação dos objetivos
- Escolha do algoritmo
- Interpretação dos resultados

31

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."



32

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

Refinamento Sucessivo

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

- Conhecimento deve ser verdadeiro
- Adequado ao contexto da aplicação (Diagnóstico de Doenças x Livros Amazon)

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

- Descobertas óbvias não interessam
- Conhecimento gerado deve ser novo

35

Descoberta de Conhecimento em BD

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."

- Conhecimento deve proporcionar benefícios

36

Caracterização do Processo

- Aplicação de KDD é dividida em 3 componentes
  - Problema a ser submetido
  - Recursos disponíveis para a solução do problema
  - Resultados obtidos através do uso dos recursos

Caracterização do Processo

- Problema
  - Conjunto de Dados
    - Aspecto Intensional – estrutura dos dados
    - Aspecto Extensional – Casos ou registros
  - Especialista no Domínio
    - Representa pessoa que conhece o assunto
  - Objetivos da Aplicação
    - Características esperadas do modelo
    - Exemplo: precisão mínima de 85% ao conceder crédito
    - Podem não estar muito claros no início do processo

Caracterização do Processo

- Recursos Disponíveis
  - Especialista em KDD
  - Ferramenta de KDD
    - Ambiente de Mineração de
    - Algoritmos Isolados
  - Plataforma Computaciona
    - Hardware
    - Capacidade de Processamento
    - Memória

•Dev
- **• Identificar e utilizar conhecimento *a priori sobre o problema***
- **• Escolher ferramentas e métodos**
- **• Direcionar as ações do processo**
- **• Conduzir a avaliação dos resultados**

39

Caracterização do Processo

- Resultados Obtidos
  - Modelo de Conhecimento
    - Deve ser avaliado com relação ao cumprimento das expectativas definidas nos objetivos
    - Usado para comparações
  - Históricos
    - Como os modelos de conhecimento foram gerados
    - Melhor controle do processo
    - Permitem análise e revisão das ações realizadas

40

Macroobjetivos e Orientações

- Aplicação de KDD pode ser classificada em duas dimensões
  - Orientação das Ações
    - Validação de Hipóteses Postuladas
    - Descoberta de Conhecimento
  - Macroobjetivos
    - Predição – permite fazer previsão a partir de históricos
    - Descrição – permite descrever o conhecimento existente na base

41

Aplicações

- **Bancária (aprovação de crédito),**
- **Ciências e medicina (descoberta de hipóteses, diagnóstico, classificação, predição),**
- **Comerciais (segmentação, localização de consumidores, identificação de hábitos de consumo),**
- **Engenharia (simulação e análise, reconhecimento de padrões, processamento de sinais e planejamento),**
- **Financeira (apoio para investimentos, controle de carteira de ações),**
- **Gerencial (tomadas de decisão, gerenciamento de documentos),**
- ***Internet(ferramentas de busca, navegação, extração de dados),***
- ***Manufatura(modelagem e controle de processos, controle de qualidade, alocação de recursos),***
- ***Segurança(detecção de bombas, icebergs e fraudes) etc.*** Análise de Churn

42

Bibliografia do Curso

- Passos, Emmanuel; Goldschmidt, Ronaldo: "Data Mining – Um Guia Prático", Editora Campus
- Witten, Ian H.; Frank, Eibe: "Data Mining", Elsevier
- AMARAL,F.C.N. *Data Mining: Técnicas e Aplicações para o Marketing Direto*. São Paulo: Editora Berkeley, 2001.
- BUSSAB,W.O. , MORETTIN,P.A. *Estatística Básica*. 5.ed. São Paulo: Editora Saraiva, 2002.
- BUSSAB,W.O. , MIAZAKI,É.S. ANDRADE,D.F. *Introdução à Análise de Agrupamentos*. São Paulo: 9° Simpósio Nacional de Probabilidade e Estatística, 1990.
- BERRY,M.J.A., LINOFF,G. *Data Mining Techniques For Marketing, Sales and Customer Support*. 2ª. ed. New York: John Wiley & Sons, Inc., 2004.
- CARVALHO,L.A.V. *Datamining A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração*. São Paulo: Editora Érica, 2001.
- DINIZ,C.A.R. , NETOF.L. *Data Mining: Uma Introdução*. São Paulo: XIV Simpósio Nacional de Probabilidade e Estatística. IME-USP, 2000.
- FERNADEZ,G. *Data Mining Using SAS Applications*. New York: Editora Chapman & Hall/CRC, 2003.
- HAN, J. , KAMBER, M.. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers, 2001.
- KLÖSGEN,W. , ZYTKOW, J.M.. *Handbook of Data Mining and Knowledge Discovery*. New York: Oxford University Press, Inc., 2002.
- MANLY,B.F.J. *Multivariate Statistical Methods: a Primer*. 3.ed. New York: Chapman & Hall, 2005.
- Oliveira, I. (2003). Correlated Data in Multivariate Analysis. *Ph.D Thesis*, University of Aberdeen.
- Jolliffe, I.T. (2002). *Principal Component Analysis. Springer Verlag – New* York Inc.
- Christensen, R. (1997). *Log-linear models and logistic regression*. NY: Springer-Verlag.
- P. McCullagh and J.A. Nelder, Generalized Linear Models 2nd edition, Chapman & Hall 1997

Software

- WEKA
- http://www.cs.waikato.ac.nz/ml/weka/