SCC0173 – Mineração de Dados Biológicos

Aula 1 – Conceituação

Prof. Ricardo J. G. B. Campello

Estagiário PAE: Pablo A. Jaskowiak

SCC / ICMC / USP

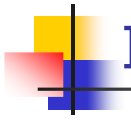
Créditos

- O material a seguir consiste de adaptações e extensões dos originais gentilmente cedidos:
  - pelo Prof. André C. P. L. F. de Carvalho
    - e seu estagiário PAE Ricardo Cerri
  - pelo Prof. João Luis Garcia Rosa

2

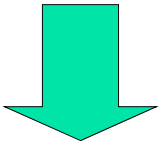
Resumo

- Introdução
- Descoberta de Conhecimento em Bases de Dados (KDD)
- Etapas de KDD
- Mineração de Dados
- Aplicações



Introdução

- Avanços recentes nas tecnologias de aquisição, transmissão e armazenamento de dados

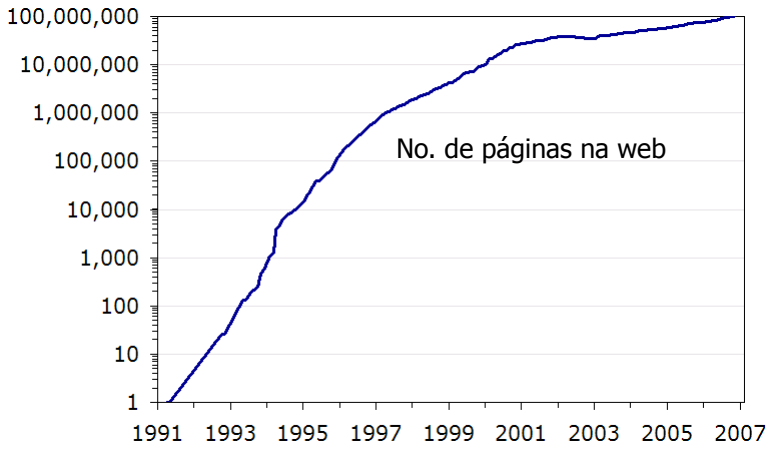


Bases de dados cada vez maiores

Introdução

- Estima-se que a quantidade de dados em Bases de Dados mundiais dobra a cada 20 meses
  - Transações bancárias
  - Utilização de cartões de crédito
  - Dados governamentais
  - Medições ambientais
  - Dados clínicos
  - Informações disponíveis na web
  - Projetos genoma

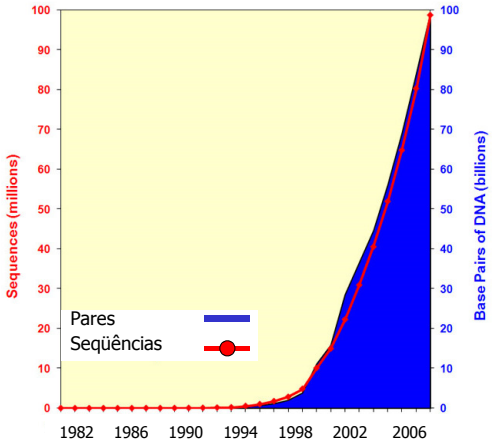
Introdução



Introdução

Crescimento do GenBank
1982-2009

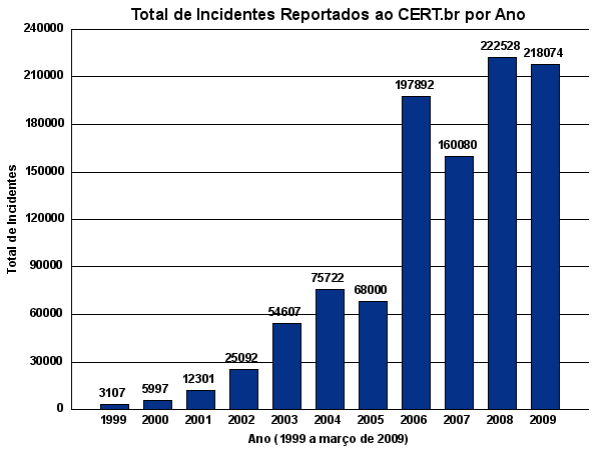
Fonte: Genbank



André Ponce de Leon F de Carvalho

Introdução

Incidentes de Segurança na Internet Brasileira

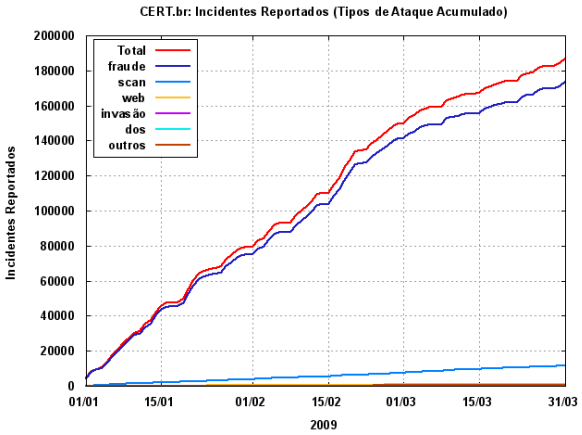


André Ponce de Leon F de Carvalho 8

Introdução

Incidentes de Segurança na Internet Brasileira

Janeiro a Março 2009



Introdução

- Grandes bases de dados não são incomuns:
  - Informação de pontos de venda, registros governamentais, registros médicos e dados de cartões de crédito, etc.
  - Instrumentos científicos podem produzir terabytes (10^{12} ou 2^{40}) e petabytes (10^{15} ou 2^{50}) a taxas de gigabytes (10^{9} ou 2^{30}) por hora
  - Capacidades de armazenamento aumentam
    - Bases de dados mais baratas e maiores, crescendo tanto em tamanho de campo e registro como em número de registros
- Limites humanos!

Introdução

- Dois exemplos:
  - Transações eletrônicas
    - BD Wal-Mart: 7.2 bilhões de diferentes transações de compra em 2008
  - Controle e monitoramento
    - BD NASA: recebe de satélites ~ 50 GB / hora

Introdução

- Bases de Dados muito grandes podem conter (esconder) dados preciosos
- Existe um interesse crescente em explorar esses dados armazenados
  - Descobrir conhecimento novo e útil
  - Suporte a decisão

Introdução

- Técnicas tradicionais de análise de dados permitem apenas consultas simples
  - P. ex., "Quantos itens de um produto em particular foram vendidos em um dado dia?"
  - Não conseguem responder consultas do tipo:
    - Que tecidos podem estar com tumor?
    - Qual a estrutura ternária de uma nova proteína
  - Técnicas mais sofisticadas, capazes de extrair **conhecimento** de grandes BD, são necessárias

Conhecimento

Compreensão ou Modelo de um determinado Domínio ou Área

Dados – Informação – Conhecimento

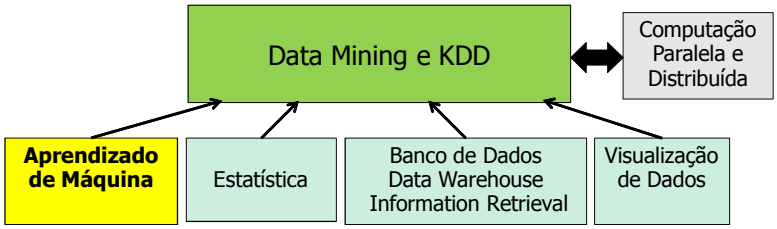
- **Dados:**
  – quantificações de um determinado evento
- **Informação:**
  – dados contextualizados
- **Conhecimento:**
  – (modelo das) relações existentes entre as informações contidas nos dados e do seu significado

KDD

- **Descoberta de Conhecimento em Bases de Dados** (KDD)
  - Extração de **conhecimento** a partir de:
    - Registros de compras em grandes supermercados
    - Registros de empréstimos financeiros
    - Registros de transações financeiras
    - Registros médicos
    - Sequências de expressão-gênica
    - ...
  - Engloba a **Mineração de Dados** (Data Mining – DM)

KDD e DM

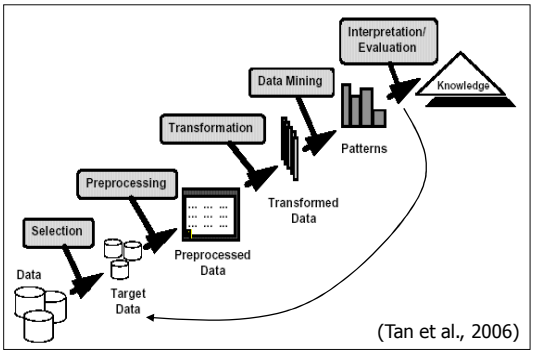
- DM se refere especificamente às ferramentas (descritivas ou preditivas) de processamento de dados para extração de conhecimento:



KDD e DM

- KDD é usualmente definido como um **ciclo** que envolve DM:
  - Formalmente, é o processo *não trivial* de identificar padrões **válidos**, **novos**, **potencialmente úteis** e **compreensíveis** em dados

Fayyad, U. M., Piaetsky-Shapiro, G. and Smyth, P., "*From Data Mining to Knowledge Discovery: An Overview*" In **Advances In Knowledge Discovery and Data Mining. AAAI/MIT Press,** p. 1–34, 1996.



(Tan et al., 2006)

KDD e DM

- DM e KDD podem ser vistos como uma evolução dos processos tradicionais de análise de BDs:

Passo evolutivo	Tecnologias existentes	Provedores de produtos	Características
Coleção de Dados (1960s)	Computadores, fitas, discos.	IBM, CDC.	Retrospectivo, fornecimento estático de dados.
Acesso aos Dados (1980s)	RDBMS, SQL, ODBC.	Oracle, IBM, Microsoft.	Retrospectivo, fornecimento dinâmico de dados no nível do registro.
Data Warehousing (1990s)	Relational Data Warehouse, OLAP, MDDB.	NCR, Business Objects, COGNOS, Hyperion.	Retrospectivo, fornecimento dinâmico de dados em vários níveis.
Data Mining (2000s)	Algoritmos avançados, Bases de Dados muito grandes.	SAS, SPSS, IBM, Oracle, NCR.	Prospectivo, informação proativa e fornecimento de conhecimento.

Slide adaptado do original do Prof. João L. G. Rosa 18

KDD e DM

- Podem tratar com terabytes, ou mais...
- Hoje, petabytes não são incomuns...
- Quão grande é um petabyte?
  - 250 bilhões de páginas de texto,
  - 20 milhões de arquivos de aço de 4 gavetas,
  - 1 torre de mais de 3.000 km de altura de 1 bilhão de disquetes:
    - O Pico da Neblina, a mais alta montanha do Brasil, tem quase 3 km
      - foto: http://www.ecotour.nl/Reizen/Expeditie%20Pico%20da%20Neblina/neb4.jpg.

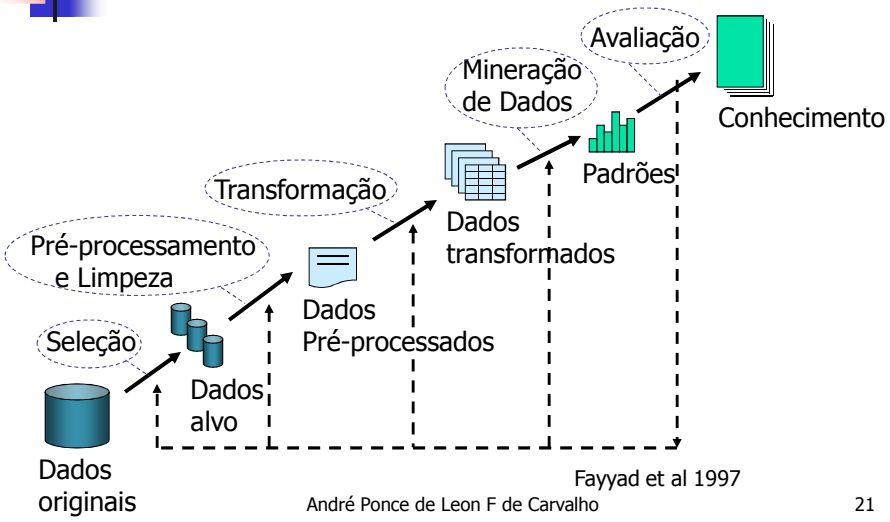
Slide adaptado do original do Prof. João L. G. Rosa 19

Tamanho dos Dados

- Tamanhos de conjuntos de dados
  - Pequeno
    - Conjunto de dados pode ser gerenciado pela ferramenta de KDD sozinha, geralmente em um arquivo e computador único
  - Médio
    - Necessária a integração do ambiente de KDD com Sistemas Gerenciadores de BDs (SGBDs), que gerenciam os dados
  - Grande
    - Quando o volume de dados é grande demais para ser gerenciado pelas ferramentas de um SGBD
    - Necessário sistemas mais sofisticados

André Ponce de Leon F de Carvalho 20

Voltando a KDD e DM...



Fayyad et al 1997

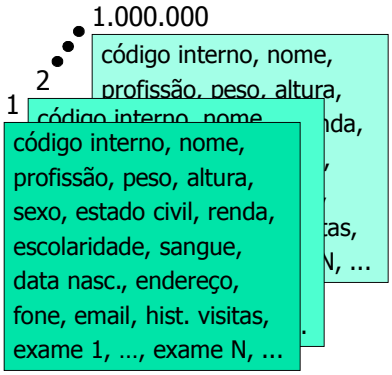
Seleção

- Começa por entender o domínio de aplicação
  - Determinar o que já é conhecido sobre o problema
  - Identificar claramente os objetivos do usuário
  - Exemplo
    - Diagnosticar paciente de acordo com conjunto de sintomas

Exemplo

- BD de um hospital
  - Composto por conjunto de registros de pacientes
  - Cada registro é composto de atributos
    - Informações pessoais
    - Sintomas

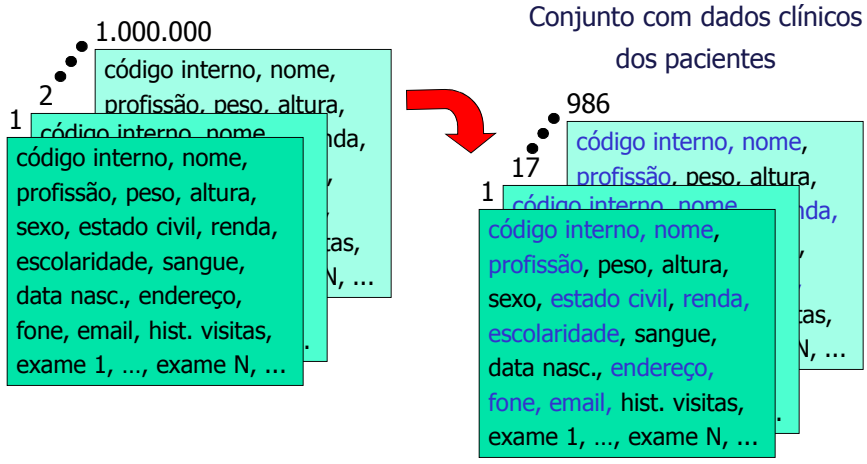
BD com registros de pacientes



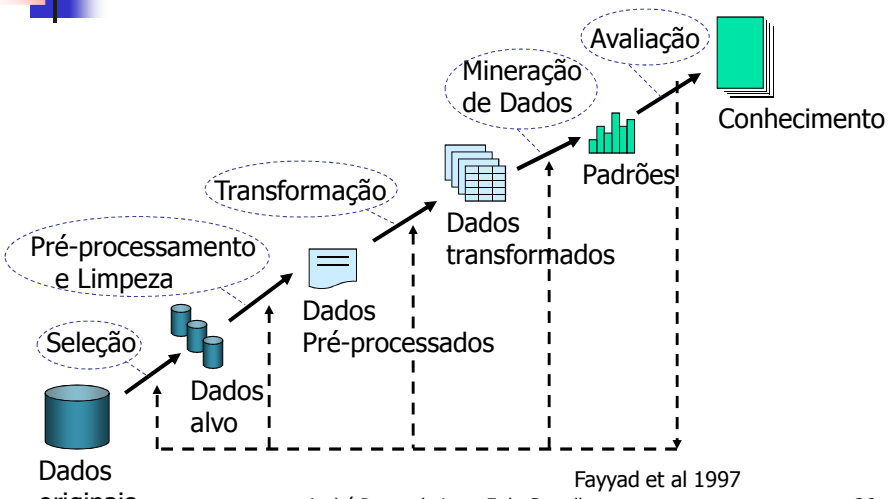
Seleção

- Toma-se então um subconjunto de interesse dentre os dados disponíveis
  - Subconjunto de registros
    - exemplos, instâncias ou objetos
  - Subconjunto de atributos
    - campos, variáveis ou características
    - considerados relevantes para o problema
      - os demais são claramente irrelevantes e descartados "manualmente"

Exemplo



Relembrando...



Pré-Processamento e Limpeza

- Melhorar a qualidade dos dados e facilitar sua posterior utilização
- As principais operações são
  - Eliminar dados duplicados
  - Lidar com valores ausentes
  - Lidar com ruído
    - p. ex. em imagens de fenômenos biológicos

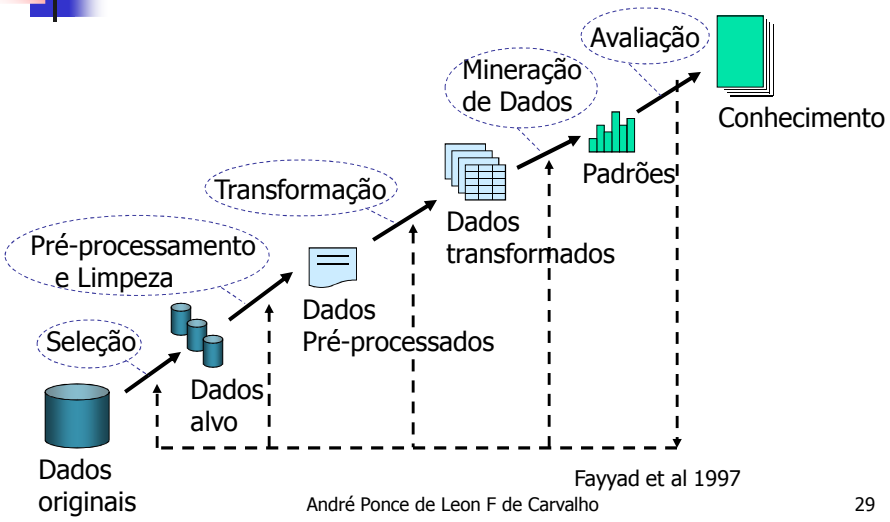
27

Transformação

- Inclui diversas operações, tais como:
  - Seleção "automática" de atributos
    - para eliminar atributos irrelevantes e/ou redundantes
  - Extração de características
    - para obter um menor no. de atributos mais relevantes a partir dos dados brutos
  - Discretizações e conversões
    - entre atributos de diferentes naturezas
  - Normalizações
    - para que os atributos ou mesmo os objetos apresentem determinadas propriedades de interesse
      - tipicamente propriedades estatísticas

28

Relembrando...

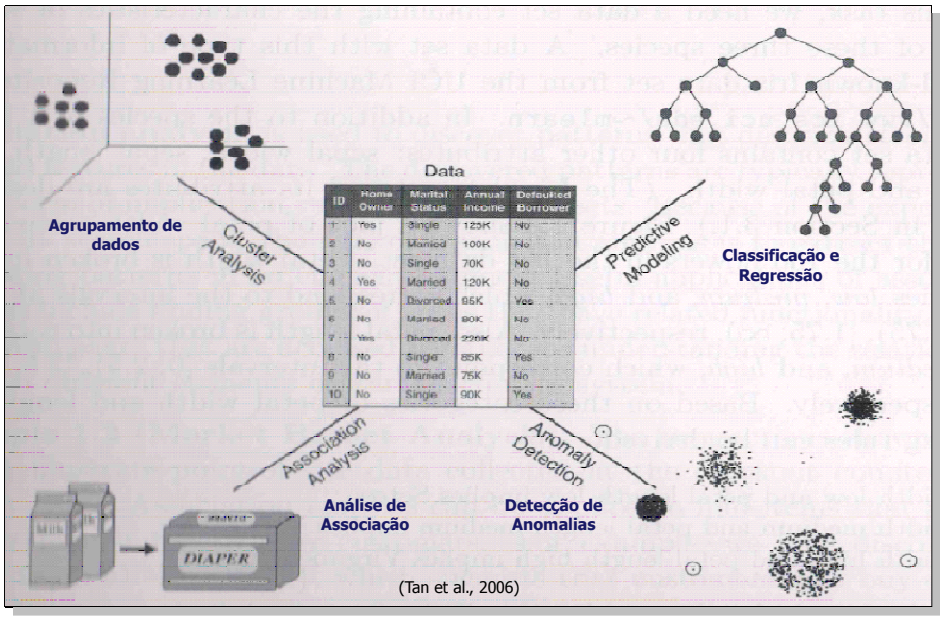


Fayyad et al 1997

Mineração de Dados (DM)

- Principal passo no processo de KDD
  - DM e KDD são freqüentemente utilizados como sinônimos
- Fronteiras da etapa de DM no processo de KDD são de difícil identificação
  - Pré-processamento e transformação de dados são freqüentemente vistos como uma parte de DM

Tarefas Fundamentais de DM



(Tan et al., 2006)

Interpretação / Avaliação

- Interpretação dos padrões obtidos na etapa de DM
  - Verifica se os padrões encontrados são:
    - válidos, novos, úteis e compreensíveis
  - Uma das etapas mais complexas
    - Em geral inclui análise estatística
    - Mas pode envolver elevado grau de subjetividade em algumas aplicações
    - Normalmente requer muito conhecimento de domínio
  - Possível retorno a qualquer uma das etapas anteriores
- Ferramentas de visualização de dados podem ter um papel de suporte importante

http://www.crisp-dm.org

CRISP-DM

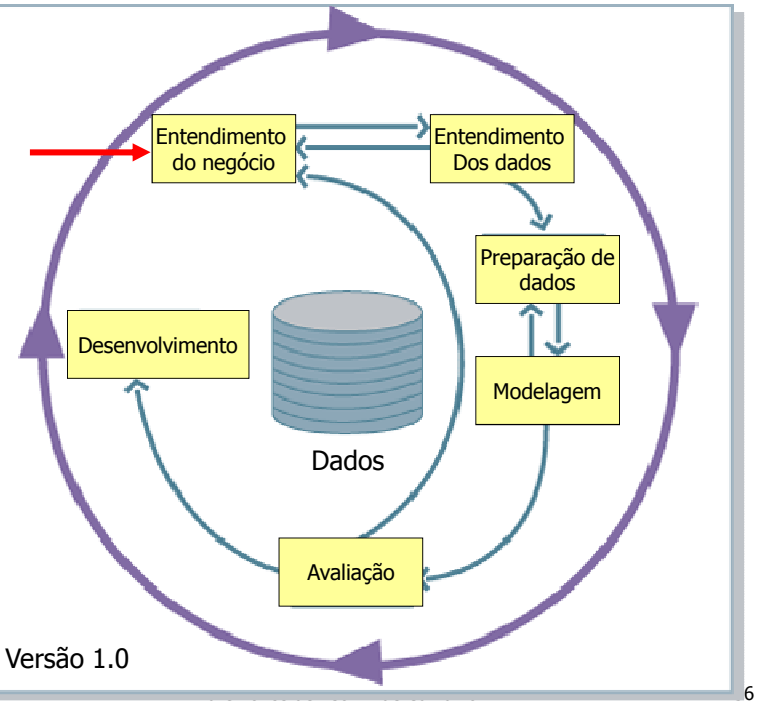
- Projeto CRISP-DM
  - *CRoss-Industry Standard Process for Data Mining*
  - Concebido em 1996 por 3 empresas:
    - Daimler-Chrysler
      - Aplicava DM em suas operações de negócios
    - SPSS
      - Prestava serviço de DM desde 1990
      - Desenvolveu primeira ferramenta comercial de DM (*Clementine*)
    - NCR
      - Tinha o propósito de adicionar valor aos seus produtos e negócios em Data Warehousing

CRISP-DM

- Projeto CRISP-DM
  - Desenvolveu um fluxo de processo para descoberta de conhecimento
    - A partir do processo anterior
      - Fayyad, Piatesky-Shapiro and Smyth
    - Em resposta a requisitos de usuários
    - Definiu e validou processos de DM utilizado em vários setores industriais

CRISP-DM

- Metodologia torna os projetos
  - Mais rápidos
  - Mais baratos
  - Mais confiáveis
  - Mais facilmente gerenciáveis
- Pode ser aplicada a pequenos projetos
- Metodologia padrão da indústria



Entendimento do Negócio

- Entender os objetivos do projeto
  - e requisitos do negócio (problema)
  - definir critérios para medir sucesso
- Converter o entendimento em definição de um problema de DM
- Traçar um planejamento preliminar para atingir objetivos

Entendimento dos Dados

- Tem início com uma coleta dos dados
- Segue com atividades para:
  - Familiarizar-se com os dados
    - Calcular estatísticas básicas
  - Explorar os dados
    - Investigar atributos
    - Identificar problemas de qualidade nos dados
  - Realizar descobertas iniciais sobre os dados
  - Detectar subconjuntos interessantes

Preparação dos Dados

- Cobre todas as atividades necessárias para construir o conjunto de dados final
  - Provavelmente executadas várias vezes
    - Sem uma ordem pré-definida
  - Inclui:
    - Seleção de tabelas, instâncias e atributos
    - Limpeza de dados
    - Transformação de dados

Modelagem

- Escolha e aplicação de diferentes técnicas
  - bem como o ajuste de seus parâmetros
- Nota:
  - Existem várias técnicas para o mesmo tipo de problema de DM...
  - Algumas têm necessidades específicas para o formato dos dados
    - Freqüentemente é necessário voltar à fase de preparação de dados

Avaliação

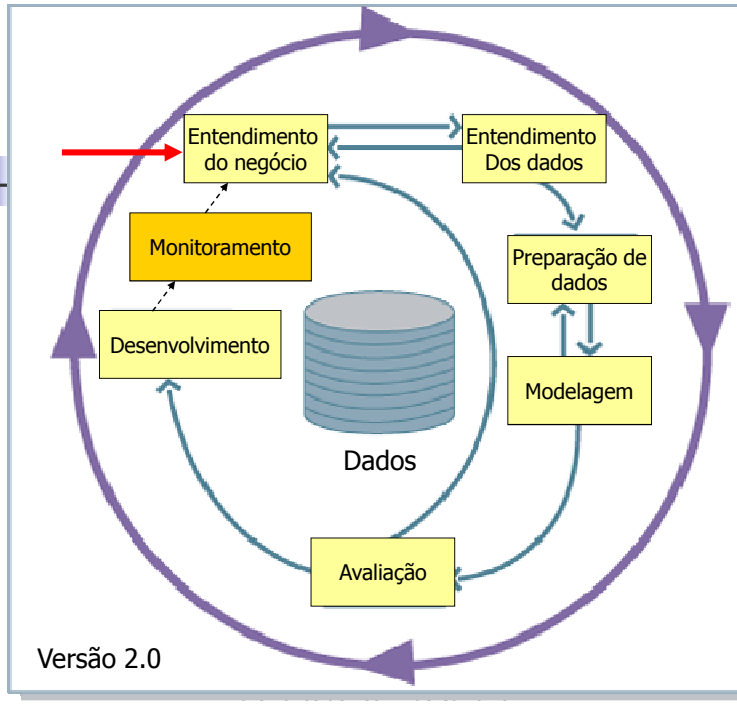
- Protótipo do modelo estará construído
- Antes do seu desenvolvimento final:
  - É importante avaliá-lo e revisar os passos executados para sua construção
    - Para ter certeza de que atende adequadamente aos objetivos do problema
    - Verificar se algum aspecto importante não foi suficientemente considerado
- Decide se os resultados da DM serão usados

Desenvolvimento

- Modelo em geral não atende às necessidades dos usuários finais, por diferentes razões, p. ex.:
  - Conhecimento precisa ser organizado e apresentado em uma forma que o usuário possa utilizá-lo
    - interfaces, ferramentas de geração de relatórios, etc.
    - usuário muitas vezes não é analista de DM !
  - Conhecimento em geral precisa ser explorado sob diferentes perspectivas por diferentes usuários e/ou em diferentes instantes de tempo
    - ferramentas de análise exploratória
  - ...

CRISP-DM

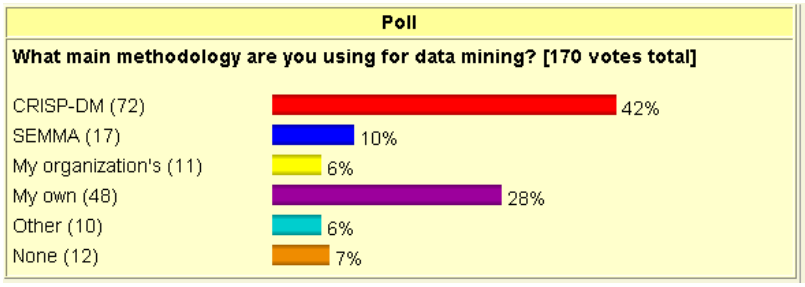
- Versão 2.0 em elaboração
  - *Special Interest Group* está sendo consultado
  - Possíveis mudanças
    - Divisão da fase de preparação de dados
    - Métodos de avaliação preliminares dentro da fase de modelagem
    - Inclusão de fase de monitoramento
    - ...



http://www.kdnuggets.com

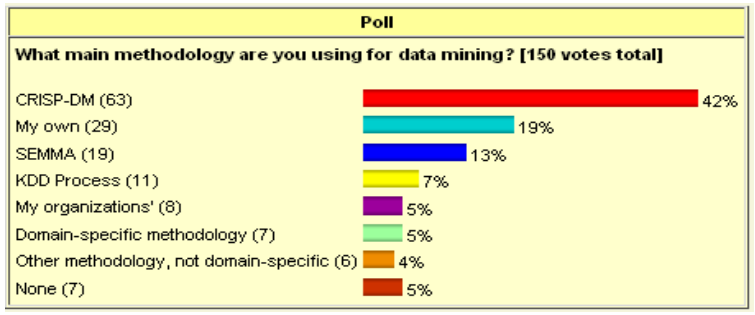
Pesquisas KDnuggets

- Pesquisa realizada em 2004



Pesquisas KDnuggets

- Pesquisa realizada em 2007



Aplicações

- Número crescente de aplicações
  - Ciência e Medicina: descoberta de padrões de risco, diagnóstico de pacientes, análise de genoma, ...
  - Finanças: análise de risco (p. ex. de crédito), detecção de fraudes, gerenciamento de carteiras, ...
  - Internet: algoritmos de busca, marketing na web, ...
  - Indústrias: previsão de falhas, diagnóstico de qualidade, ...
  - Marketing: segmentação de mercado, ...

Exemplos de Aplicações

- Siemens Medical
  - Ferramenta de DM para o Tratamento de Ataques cardíacos
  - Combina informações médicas de diversas fontes
  - Busca automática em registros combinados de 6 milhões de pacientes

Exemplos de Aplicações

- Siemens Medical
  - Descobriu centenas de casos onde os melhores procedimentos médicos não haviam sido seguidos
    - Mas ainda havia tempo para intervir
  - Identificou pacientes elegíveis para estudos médicos
  - Ganhou o 2005 ICDM Data Mining Practice Prize

Exemplos de Aplicações

- The Mitre Coorporation
  - Ferramenta de DM para detecção de fraudes no IR
  - Sistema de DM usa Aprendizado de Máquina e Análise Estatística para descobrir sonegações
  - Resultados
    - Encontrou casos não descobertos por auditores
    - Reduziu tempo de análise
      - 2 semanas para poucas horas (dados de 2001)
  - 2º lugar no 2005 ICDM Data Mining Practice Prize

Aplicações em Dados Biológicos

- Análise de expressão gênica
- Reconhecimento de genes
- Previsão de estruturas de proteínas
- Alinhamento de sequências
- Reconstrução de árvores filogenéticas
- ...

51

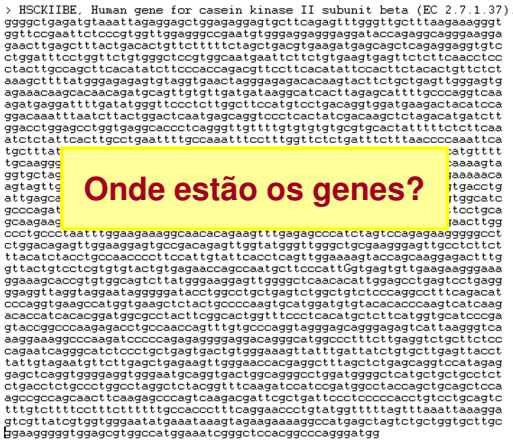
Análise de Expressão Gênica

- Medição do nível de expressão de grande quantidade de genes sob diversas condições
- Aplicações
  - Agrupar genes com expressões similares
  - Atribuir funções a genes não anotados
  - Identificar novas subclasses de doenças
  - Classificar tecidos
- Algoritmos de
  - Agrupamento
  - Classificação
  - Seleção de atributos

52

Reconhecimento de Genes

- Encontrar genes em sequências de DNA



Reconhecimento de Genes

- Encontrar genes em sequências de DNA

Reconhecimento de Genes

- Encontrar genes em sequências de DNA
- Análise em laboratório
  - Difícil e cara
- Análise computacional
  - Duas abordagens
    - Busca por sinal
    - Busca por conteúdo
  - Ambas utilizam algoritmos de Classificação

55

Previsão de Estruturas de Proteínas

- Predição da estrutura espacial de proteínas
  - Importante uma vez que sua estrutura determina ou provê indícios de sua função
- Métodos tradicionais
  - Cristalografia
- Métodos computacionais
  - Se baseiam em estruturas já conhecidas
  - Aplicação de métodos de classificação

56

Alinhamento de Sequências

- Busca por sequências similares
  - Possivelmente possuem ancestral comum
    - História evolutiva dos organismos
  - Diferenças entre sequências
    - Deleções
    - Inserções
    - Substituições
- Algoritmos
  - Programação dinâmica (BLAST)
  - Algoritmos genéticos

57

Reconstrução de Árvores Filogenéticas

- Árvores filogenéticas
  - Ancestralidade, relacionamento entre espécies ou grupos de espécies
  - Construídas com base em similaridades, por exemplo, entre sequências de bases
- Métodos computacionais
  - Método Neighbor-Joining, baseado no conceito de vizinho mais próximo.
  - Algoritmos de agrupamento hierárquicos

58

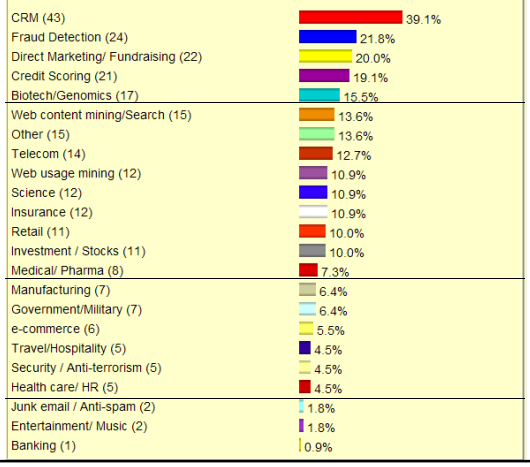
Pesquisas KDnuggets

- Aplicações de DM
- Em que indústrias / áreas você está atualmente aplicando DM
  - Fonte:
    - http://www.kdnuggets.com/polls/2006/data_mining_applications_industries.htm
    - Data: junho de 2006
    - 278 votos de 111 votantes

Pesquisas KDnuggets

Poll

Industries/fields where you applied data mining in the past 12 months [111 voters, 278 votes total]



Pesquisas KDnuggets

- Aplicações de DM
- Em que indústrias / áreas você está atualmente aplicando DM
  - Fonte:
    - http://www.kdnuggets.com/polls/2007/data_mining_applications.htm
    - Data: junho de 2007
    - 138 votantes

Pesquisas KDnuggets

Poll

Industries/fields where you applied data mining in the past 12 months [138 voters]

CRM (36)	26.1%
Banking (33)	23.9%
Direct Marketing/ Fundraising (28)	20.3%
Science (26)	18.8%
Fraud Detection (26)	18.8%
Telecom (21)	15.2%
Credit Scoring (19)	13.8%
Other (18)	13.0%
Biotech/Genomics (16)	11.6%
Web usage mining (14)	10.1%
Retail (14)	10.1%
Medical/ Pharma (13)	9.4%
Insurance (12)	8.7%
Health care/ HR (10)	7.2%
Government/Military (10)	7.2%
Financials/Lending (10)	7.2%
Web content mining/Search (9)	6.5%
Manufacturing (9)	6.5%
e-commerce (8)	5.8%
Entertainment/ Music (6)	4.3%
Social Policy/Survey analysis (5)	3.6%
Security / Anti-terrorism (5)	3.6%
Investment / Stocks (4)	2.9%
Travel/Hospitality (3)	2.2%
Junk email / Anti-spam (3)	2.2%

Pesquisas KDnuggets

- 2006 x 2007

Industry/Area	2006 Count	2007 Count	Change Y6 to Y7
Banking	1	33	3200%
Entertainment/ Music	2	6	200%
Science	12	26	117%
Health care/ HR	5	10	100%
Medical/ Pharma	8	13	63%
Junk email / Anti-spam	2	3	50%
Telecom	14	21	50%
Government/Military	7	10	43%
e-commerce	6	8	33%
Manufacturing	7	9	29%
Direct Marketing/ Fundraising	22	28	27%
Retail	11	14	27%
Other	15	18	20%
Web usage mining	12	14	17%
Fraud Detection	24	26	8%
Insurance	12	12	0%
Security / Anti-terrorism	5	5	0%
Biotech/Genomics	17	16	-6%
Credit Scoring	21	19	-10%
CRM	43	36	-16%
Travel/Hospitality	5	3	-40%
Web content mining/Search	15	9	-40%
Investment / Stocks	11	4	-64%

Pesquisas KDnuggets

- Aplicações de DM
- Em que indústrias / áreas você está atualmente aplicando DM
  - Fonte:
    - http://www.kdnuggets.com/polls/2008/data-mining-applications.htm
    - Data: dezembro de 2008
    - 107 votantes

Pesquisas KDnuggets

Poll

Industries / Fields where you applied Data Mining in 2008: [107 voters]

CRM/ consumer analytics (41)	38.3%
Banking (34)	31.8%
Fraud Detection (21)	19.6%
Finance (18)	16.8%
Direct Marketing/ Fundraising (15)	14.0%
Other (14)	13.1%
Investment / Stocks (14)	13.1%
Credit Scoring (14)	13.1%
Telecom / Cable (13)	12.1%
Retail (13)	12.1%
Advertising (13)	12.1%
Biotech/Genomics (12)	11.2%
Science (11)	10.3%
Insurance (11)	10.3%
Health care/ HR (10)	9.3%
Manufacturing (9)	8.4%
e-Commerce (8)	7.5%
Web usage mining (8)	7.5%
Social Policy/Survey analysis (8)	7.5%
Medical/ Pharma (8)	7.5%
Security / Anti-terrorism (6)	5.6%
Search / Web content mining (6)	5.6%
Government/Military (4)	3.7%
Travel / Hospitality (3)	2.8%
Junk email / Anti-spam (3)	2.8%
Entertainment/ Music (3)	2.8%
Social Networks (2)	1.9%
None (2)	1.9%

Produtos de DM

Mais Produtos

Alguns Mitos (Padhraic Smith)

- "Análise de dados pode ser completamente automatizada"
  - Julgamento humano é crítico na maioria das aplicações
  - De qualquer modo, semi-automatização é muito útil
- "Com uma quantidade massiva de dados, não é necessário estatística"
  - Grande volume leva a heterogeneidade
    - Precisa ainda mais de estatística

Pacotes e Conjuntos de Dados

- WEKA
  - http://directory.google.com/Top/Computers/Artificial_Intelligence/Machine_Learning
- R
  - http://lancet.mit.edu/ga
- Machine Learning Data Repository UC Irvine
  - http://www.ics.uci.edu/~mlearn/ML/Repository.html

Considerações Finais

- Expansão do volume de **dados**
- Necessidade de extrair **conhecimento**
- KDD é cada vez mais usado
- Cuidado com promessas exageradas
  - Sistemas Especialistas...

Perguntas