MINERAÇÃO DE DADOS E TEXTOS

SCC-230 Inteligência Artificial

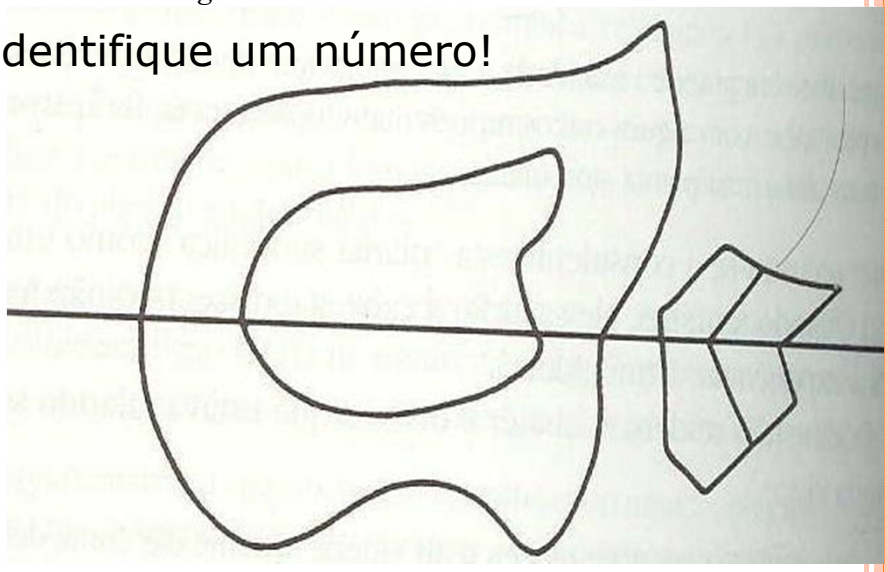
Solange Oliveira Rezende
Bruno Magalhães Nogueira
Thiago A. S. Pardo

MOTIVAÇÃO

LABIC USP - São Carlos

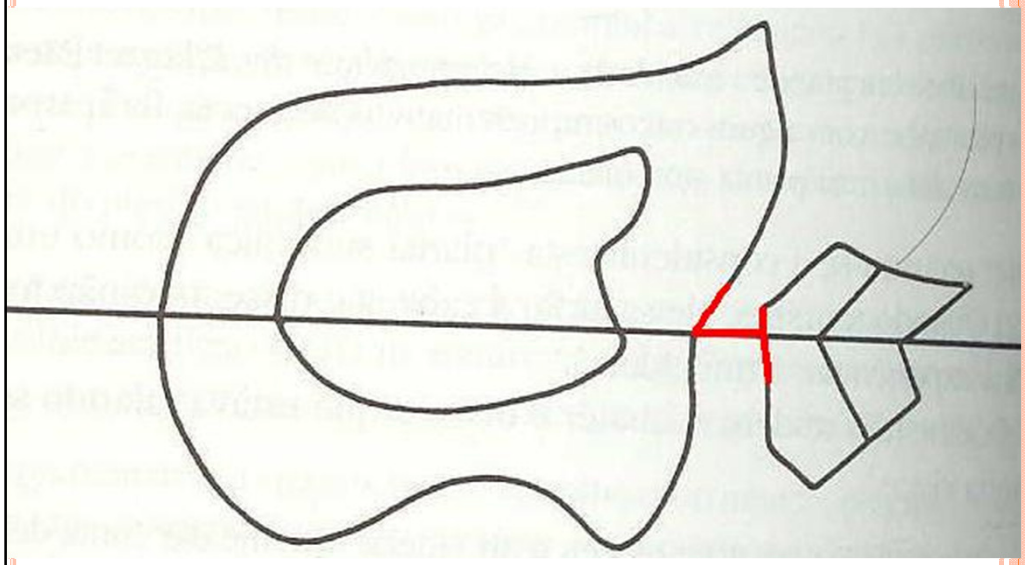
Observe a imagem...

Identifique um número!



2

MOTIVAÇÃO



3

MOTIVAÇÃO

Cassino Harrah's
(Guizzo, 2001)
16 milhões de clientes!

Qual o perfil de cliente proporciona maior lucratividade?

- Apostadores que gastam entre US$ 100 a 500:
  - ✓ Representam 30% da clientela
  - ✓ Contribuem com 80% das receitas
- Estratégias de marketing para este "filão" mais rentável dobrou o faturamento

4

MOTIVAÇÃO

NIKE

WAL MART

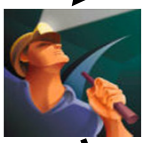
WAL★MART

5

MOTIVAÇÃO

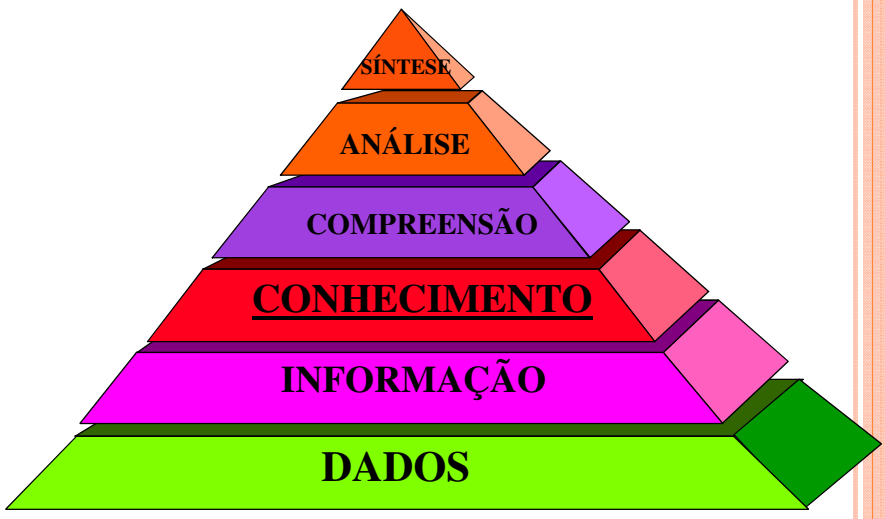
- Os sistemas computacionais armazenam quantidades cada vez maiores de dados.
  - Esse volume de dados é uma valiosa fonte de conhecimento.
- A quantidade e complexidade dos dados impossibilitam a exploração manual desse conhecimento.

Necessidade de técnicas automáticas para extrair padrões dos dados armazenados.



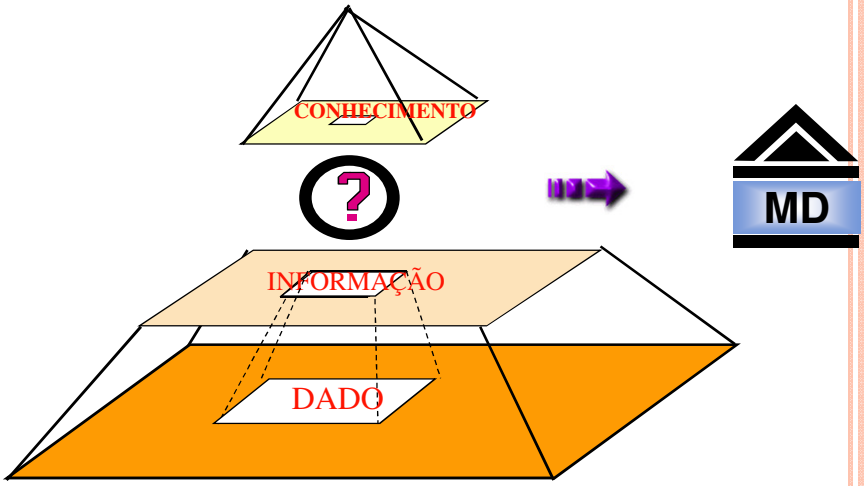
6

De Dados a Manipulação de Conhecimento: uma Estrutura



7

Por que Tecnologias como Mineração de Dados?

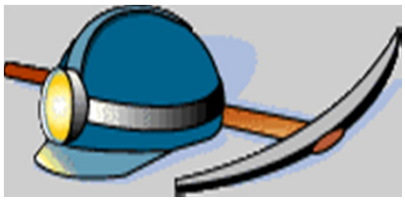


8

Parte 1: Mineração de Dados

DEFINIÇÕES

- Mineração de Dados (MD) refere-se ao processo de extrair conhecimento de bases de dados, ou seja, trabalhar com grandes quantidades de dados com o objetivo de extrair significado e descobrir novos conhecimentos.



LABIC
USP - São Carlos

Metodologias e Abordagens de Mineração de Dados

TOP DOWN

INICIAR COM HIPÓTESES E VALIDAR AS MESMAS

AS HIPÓTESES PODEM SER CONSTITUÍDAS INICIALMENTE A PARTIR DA ABORDAGEM *BOTTOM UP* OU A PARTIR DE ALGUM CONHECIMENTO DO "MUNDO REAL"

SE A HIPÓTESE NÃO FOR SATISFEITA, REVISÁ-LA

BOTTOM UP

ANALISAR OS DADOS E EXTRAIR PADRÕES

SUPERVISIONADO:

TEM-SE ALGUMA IDÉIA DO QUE ESTÁ PROCURANDO

NÃO SUPERVISIONADO:

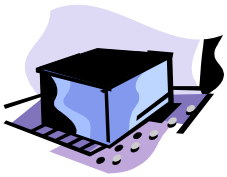
NÃO SE TEM IDÉIA DO QUE ESTÁ PROCURANDO

(Rezende et al., 2003)

11

LABIC
USP - São Carlos

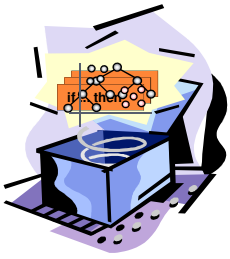
Dados



Conhecimento

Caixa preta

Dados



Conhecimento

Caixa semi-transparente
Orientada à Conhecimento

12

O Processo de Mineração de Dados

(Rezende et al., 2003)

13

Identificação do Problema

14

Identificação do Problema

- Estudo do domínio de aplicação
- Definição e identificação dos objetivos

- Quais as principais metas do processo???
- Quais critérios de desempenho são importantes?
- O conhecimento extraído deve ser compreensível a seres humanos ou o modelo do tipo caixa-preta é apropriado?
- Qual deve ser a relação entre simplicidade e precisão do conhecimento extraído?

15

Pré-Processamento

16

PRÉ-PROCESSAMENTO

- Transformação nos dados para deixá-los adequados para a etapa de Extração de Padrões
  - Extração e Integração
  - Transformação
  - Limpeza
  - Redução de Dados

17

PRÉ-PROCESSAMENTO - EXTRAÇÃO E INTEGRAÇÃO

- Os dados podem estar em diferentes formatos, como arquivos texto, arquivos no formato MS EXCEL, banco de dados relacionais, DataWarehouse.

- É necessário a unificação formando uma única fonte de dados

	X_1	X_2	...	X_m	Y
E_1	x_{11}	x_{12}	...	x_{1m}	y_1
E_2	x_{21}	x_{22}	...	x_{2m}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
E_1	x_{n1}	x_{n2}	...	x_{nm}	y_n

18

PRÉ-PROCESSAMENTO - TRANSFORMAÇÃO

- Adequação aos algoritmos de Extração de Padrões
  - Resumo
  - Transformação de tipo
  - Normalização de atributos contínuos
- Podem ser muito importantes em alguns domínios, como em aplicações que envolvem séries temporais como predições no mercado financeiro

19

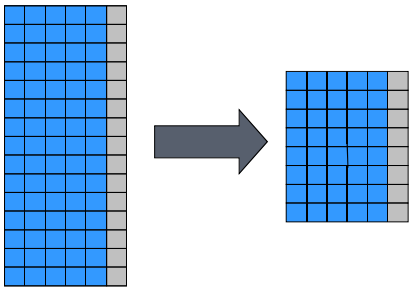
PRÉ-PROCESSAMENTO - LIMPEZA

- Dados podem apresentar problemas provenientes da coleta (digitação ou leitura por sensores)
- Qualidade é muito importante
  - Utilizar conhecimento do domínio
  - Decisão da estratégia de tratamento de atributos incompletos, remover ruídos

20

Pré-Processamento
- Redução de Dados

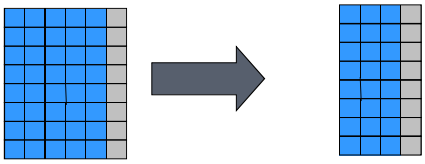
- Limitações de espaço em memória, tempo de processamento
- A redução pode ser realizada de três formas:
  - **Número de exemplos**



21

Pré-Processamento
- Redução de Dados

- Limitações de espaço em memória, tempo de processamento
- A redução pode ser realizada de três formas:
  - Número de exemplos
  - **Número de atributos**

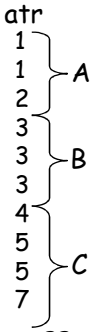


22

PRÉ-PROCESSAMENTO
- REDUÇÃO DE DADOS

- Limitações de espaço em memória, tempo de processamento
- A redução pode ser realizada de três formas:
  - Número de exemplos
  - Número de atributos
  - **Número de valores**
    - *Discretização*

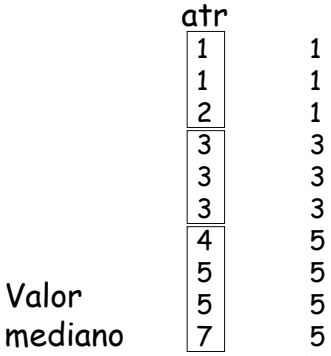
A se atr < 2,5
B se 2,5 ≤ atr < 3,5
C se 3,5 ≤ atr



23

PRÉ-PROCESSAMENTO
- REDUÇÃO DE DADOS

- Limitações de espaço em memória, tempo de processamento
- A redução pode ser realizada de três formas:
  - Número de exemplos
  - Número de atributos
  - **Número de valores**
    - Discretização
    - *Suavização*



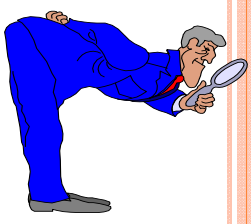
24

Extração de Padrões

25

Extração de Padrões

- Etapa é direcionada ao cumprimento dos objetivos identificados na fase de identificação do problema
- Processo iterativo
  - Escolha da Atividade e da Tarefa
  - Escolha do Algoritmo
  - Extração dos Padrões



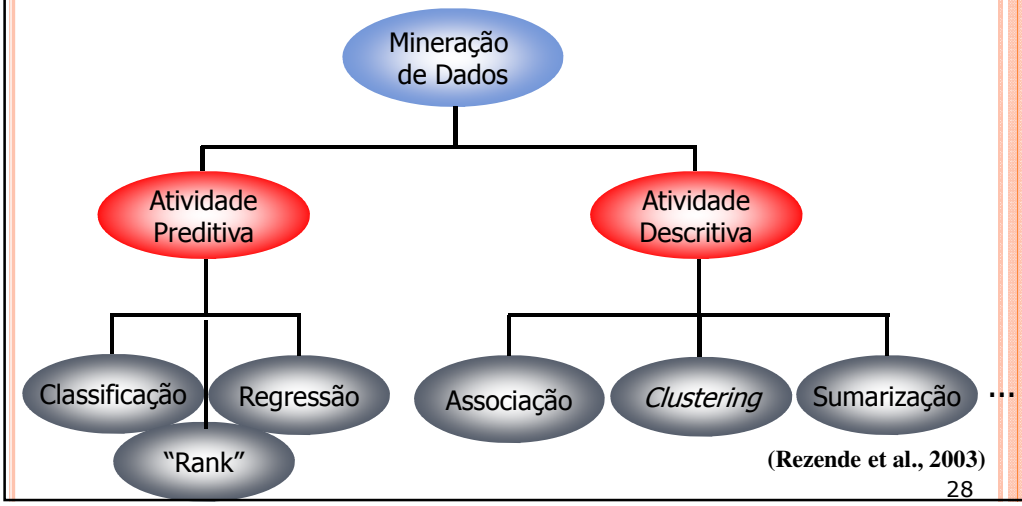
26

Extração de Padrões - Escolha da Atividade e da Tarefa

- Deve ser feita de acordo com os objetivos desejáveis para a solução a ser encontrada
- Atividades podem ser agrupadas em:
  - Atividades Preditivas
    - corresponde ao aprendizado supervisionado
  - Atividades Descritivas
    - corresponde ao aprendizado não-supervisionado

27

Extração de Padrões - Escolha da Atividade e da Tarefa



(Rezende et al., 2003)

28

EXTRAÇÃO DE PADRÕES
- ESCOLHA DO ALGORITMO

- Para efetuar a busca de padrões podem ser utilizados Algoritmos de Aprendizado de Máquina, ou outros.
- A escolha de um algoritmo é vista como um processo analítico, pois nenhum deles tem desempenho ótimo em todos os domínios de aplicação.

29

EXTRAÇÃO DE PADRÕES
- ESCOLHA DO ALGORITMO (CONT)

- Um fator relacionado com a configuração dos parâmetros dos algoritmos é a complexidade da solução a ser buscada
- Vários algoritmos estão disponíveis para cada atividade

- Representação do Conhecimento
  - Árvores de Decisão
  - Regras de Produção
  - Redes Neurais Artificiais

30

EXTRAÇÃO DE PADRÕES
- EXECUÇÃO

- Aplicação do algoritmo escolhido
- Geralmente, os algoritmos são executados diversas vezes. Alguns casos em que isso ocorre são:
  - Estimativa da taxa de erro
    - Exemplos: *cross-validation*
  - Combinação de preditores
    - Obter um preditor mais preciso

31

PÓS-PROCESSAMENTO

32

PÓS-PROCESSAMENTO

- Avaliação do conhecimento extraído
  - *O conhecimento extraído representa o conhecimento do especialista?*
  - *De que maneira o conhecimento do especialista difere do conhecimento extraído?*
  - *Em que parte o conhecimento do especialista está correto?*

33

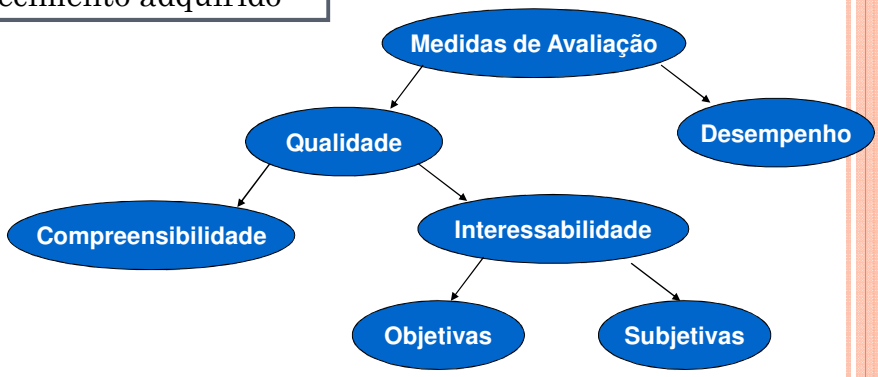
PÓS-PROCESSAMENTO - AVALIAÇÃO DO CONHECIMENTO

- Pode-se ter uma quantidade enorme de padrões que podem não ser importantes, relevantes ou interessantes aos usuários
- Não é muito interessante fornecer uma quantidade grande de padrões ao usuário, para ser avaliado
  - Desenvolver técnicas de apoio para fornecer padrões mais interessantes

34

PÓS-PROCESSAMENTO
- MEDIDAS DE AVALIAÇÃO

Existem diversas medidas para auxiliar o usuário no entendimento e na utilização do conhecimento adquirido



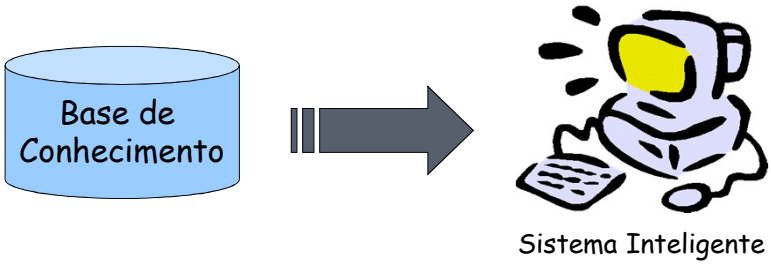
35

UTILIZAÇÃO DO CONHECIMENTO

36

UTILIZAÇÃO DO CONHECIMENTO

- Incorporando-o a um sistema inteligente
  - Apoio à tomada de decisão
  - Relatar às pessoas interessadas



37

DISPONIBILIZAÇÃO DO CONHECIMENTO

- Após a análise do conhecimento, se os resultados não forem satisfatórios, o processo de extração pode ser reiniciado com o objetivo de se obter melhores resultados

- No final do processo de MD é interessante que todo o conhecimento adquirido seja disponibilizado em um ambiente adequado para facilitar sua exploração, interpretação e utilização

38

Parte 2: Mineração de Textos

MINERAÇÃO DE TEXTOS

- Mineração de Textos trata da descoberta de conhecimento útil em grandes **coleções de textos** em meio digital
- Dados não estruturados ou semi-estruturados

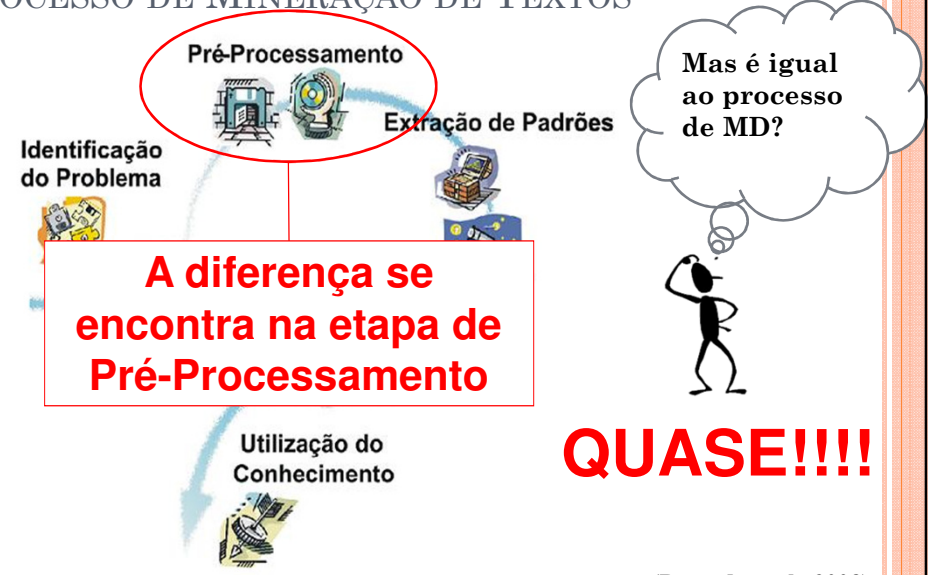
Aplicações da Mineração de Textos

- Organização das coleções textuais em bases de dados
- Inteligência competitiva
- *Matching* de documentos
- Categorização
- Filtros para e-mail (anti-spams)
- Máquinas de busca mais inteligentes
- Extração de informação (auxilia o reconhecimento de padrões)
- "Customização de jornal"...
- .....

~80% da informação é textual

41

Processo de Mineração de Textos



A diferença se encontra na etapa de Pré-Processamento

QUASE!!!!

(Rezende et al., 2003)

42

DIFERENÇA ENTRE OS PROCESSOS DE MD E MT

- **Entrada para o processo de MD**

Nome	Idade	Renda	Crédito
José	<=30	Baixa	Ruim
João	<=30	Baixa	Bom
Maria	31..40	Alta	Bom
Mario	>40	Média	Ruim
Marcos	>40	Média	Ruim
Tiago	31..40	Alta	Bom
...	...	...	...

Tabela atributo-valor

- **Entrada para o processo de MT**

Coleção de Documentos

Como fazer essa transformação?

DIFERENÇA ENTRE OS PROCESSOS DE MD E MT

- Etapa de Pré-processamento da MT tem uma tarefa adicional se comparada à mesma etapa da MD: **Estruturação dos documentos**

- Inclui três grandes sub-etapas:
  - Adequação da coleção de documentos;
  - Geração de atributos e redução do número de atributos;
  - Estruturação em formato manipulável por algoritmos de extração de conhecimento.

PRÉ-PROCESSAMENTO EM MT

- Adequação da coleção de documentos
  - Verificar se a coleção é suficiente e adequada aos objetivos do processo
    - Eliminação da repetição de documentos;
    - Balanceamento da coleção por reamostragem;
    - Redução da quantidade de documentos;
    - Verificação de estrutura prévia nos documentos;
    - Separação da coleção por tamanho dos documentos;
    - Separação da coleção por idioma dos documentos.
  - O analista deve verificar, neste ponto, se os documentos disponíveis são suficientes
    - Caso não sejam, a coleção deve ser completada.

45

PRÉ-PROCESSAMENTO EM MT

- Geração de atributos e redução do número de atributos
  - Cada termo presente na coleção é candidato a atributo;
  - Além disso, é possível considerar combinações de termos subsequentes (n-gramas) como atributos;
  - Número de termos gerados é, geralmente, muito grande, excedendo a quantidade de documentos em mais de uma ordem de magnitude
    - Representações esparsas da coleção;
    - Impacto negativo na eficiência de algoritmos de aprendizado.
  - Necessidade de gerar termos representativos e selecionar os mais importantes aos objetivos da aplicação.

46

PRÉ-PROCESSAMENTO EM MT

- Geração de atributos simples
  - Busca obter termos que sejam semanticamente significativos;
  - Em um primeiro momento, desconsidera-se da coleção termos que nada acrescentam ao domínio, denominados ***stopwords***
    - Preposições, artigos, interjeições, etc;
    - *Stopwords* de domínio – palavras que, especificamente para aquele domínio, devem ser desconsideradas;
  - Posteriormente, busca-se identificar palavras similares quanto ao seu significado
    - Variações morfológicas: *stemming*, lematização, substantivação, etc;
    - Sinônimos: *thesaurus* ou dicionários.

47

PRÉ-PROCESSAMENTO EM MT

- Geração de atributos compostos
  - A partir dos termos simples obtidos, busca-se gerar combinações de termos que expressem um conceito único;
  - Geralmente, usa-se alguma medida estatística que aponte a representatividade dos termos gerados
    - Ex: Suponha o bigrama "inteligência artificial"

	Artificial	Outros termos
Inteligência	# Inteligência_Artificial	# Inteligência_X
Outros termos	# X_Artificial	# X_Y

    - Utilizando algum teste estatístico, descarta-se os irrelevantes
      - Ex: Teste de máxima verossimilhança – lida bem com dados esparsos.

48

PRÉ-PROCESSAMENTO EM MT

- Redução do número de atributos
  - Mesmo com uma geração mais apurada, o número de atributos é geralmente muito grande;
  - Há a necessidade de reduzir o número de atributos presentes na base sem, no entanto, afetar a qualidade do resultado final do processo;
  - **Extração** x **Seleção** de Atributos.

49

PRÉ-PROCESSAMENTO EM MT

- Extração de Atributos
  - Criação de um novo conjunto de atributos com menor dimensionalidade;
  - Uso de uma função de mapeamento entre as representações;
  - Atributos obtidos são combinações dos originais;
  - Principal desvantagem: atributos gerados não mantêm correlação explícita com a configuração original do problema
    - Modelos gerados são mais difíceis de se interpretar;
  - Exemplos de técnicas: *Principal Component Analysis* (PCA) e *Latent Semantic Analysis* (LSA).

50

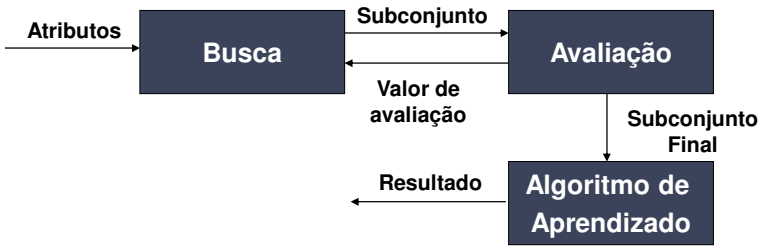
PRÉ-PROCESSAMENTO EM MT

- Seleção de atributos
  - Consiste em obter um subconjunto de atributos a partir do conjunto original, seguindo alguns critérios;
  - Mantém a relação física com o problema real;
  - Existem dois *frameworks* básicos para selecionar atributos: filtros e *wrappers*.

51

PRÉ-PROCESSAMENTO EM MT

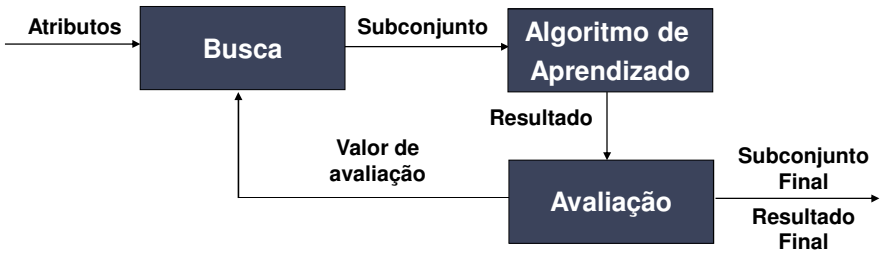
- Filtros (*filtering*)
  - Pré-selecionam os atributos e então aplicam o subconjunto ao algoritmo de aprendizado.



52

PRÉ-PROCESSAMENTO EM MT

- Wrappers
  - Incorporam o algoritmo de aprendizado no processo de busca e seleção.



53

PRÉ-PROCESSAMENTO EM MT

- Escolha do método de redução do número de atributos depende da existência ou não de rótulo nos dados
  - Dados rotulados: métodos **supervisionados** (Ganho de Informação, Informação Mútua, Chi Quadrado, etc.);
  - Dados não-rotulados: métodos **não-supervisionados** (Cortes de Luhn, Cortes de Salton, Variância do Termo, Contribuição do Termo, etc.).
- **Dados não-rotulados**: problema da avaliação dos subconjuntos de atributos
  - Difícil estabelecer uma medida que quantifique o quão bom é um subconjunto de atributos;
- **Dados rotulados**: avaliação por medidas como erro e acurácia de classificadores.

54

Pré-processamento em MT

- Estruturação da coleção em formato manipulável por algoritmos de extração de conhecimento
  - Geralmente, usa-se formato *bag-of-words*
    - Tabela atributo-valor;
    - Linhas: documentos;
    - Colunas: termos;
    - Células internas: medida de correlação entre um documento e um termo;
      - Binária: 1 caso termo ocorra no documento, 0 em caso contrário;
      - *Term Frequency* (TF): frequência absoluta do termo no documento;
      - *Term Freqüency - Inverse Document Frequency* (TFIDF): frequência absoluta do termo no documento, ponderada pelo inverso do número de documentos em que o termo ocorre.

55

Pré-Processamento em MT: Resultado

Doc	Termo1	Termo2	Termo n
Doc1	freq11	freq21	freqn1
Doc2	freq12	freq22	freqn2
Doc3	freq13	freq23	freqn3
Doc4	freq14	freq24	freqn4
Doc5	freq15	freq25	freqn5
Doc6	freq16	freq26	freqn6
...	...	...	...

- Com a tabela atributo-valor estabelecida, o restante do processo é idêntico ao processo de Mineração de Dados!

56

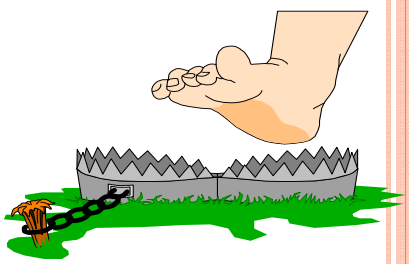
Considerações Finais

- MD é muito útil quando há dados disponíveis
- Um dos grandes problemas de MD está relacionado com a utilização/criação dos algoritmos para grandes volumes de dados
- A presença de especialistas é muito importante no processo MD
  - Se o custo da descoberta é maior que o ganho, o esforço pode não justificar!

57

Considerações Finais (cont)

- Alguns Problemas em Mineração de Dados
  - Falta de informação e buracos na sequência da informação
  - Em bases dinâmicas as trocas nos registros (tamanho, tipo, etc.) são comuns
  - Incerteza nos dados
  - Semântica embutida no dados



58