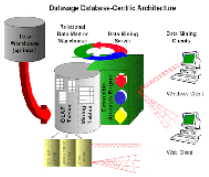
Robert Groth

Extracção de Conhecimento em Bases de Dados (ECBD ou KDD)

Usama Fayyad et al



Extracção de Conhecimento em Bases de Dados (ECBD ou KDD)

- Definição de ECBD / Data Mining
- Breve história do Data Mining
- Introdução ao Data Mining
- Tecnologias de Suporte ao Data Mining
- Fases do processo de ECBD
- Operações de Data Mining
- Métodos e Algoritmos de Data Mining
  - Soluções distância (K-vizinhos ,mais próximos e associações)
  - Naïve-Bayes
  - Arvores de decisão
  - Regras de associação
  - Redes neuronais,
  - Algoritmos genéticos.
  - Combinação de múltiplos métodos de predição.
- Alguns prós e contras das tecnologias mais comuns; ferramentas mais relevantes e suas características

Extracção de Conhecimento em Bases de Dados (ECBD ou KDD)

- O **ECBD** ou ***KDD*** é muitas vezes denominado de apenas Data Mining, ainda que, este seja, mais propriamente, uma das fases do processo *(KDD conference, 1995, Montreal)*.

Relação do ECBD com outras ferramentas de exploração de informação:

- Com ferramentas até agora descritas (capítulo anterior), poder-se-á responder a questões como:
  - **as vendas do produto X cresceram em Novembro?**
  - **as vendas do produto X diminuem quando há uma promoção do produto Y?**
- Com ferramentas no domínio do ECBD/Data Mining, poderemos colocar a questão:
  - **Quais são os factores que determinam as vendas do produto X?**

DQ/Reporting e OLAP x Data Mining

Relembrando o que já atrás foi focado (capítulo 1):

- Com as ferramentas tradicionais, o analista coloca uma questão, ou suposição ou talvez só uma inclinação e explora os dados. Cria um modelo, passo-a-passo, trabalhando para provar ou negar uma teoria.
- É da responsabilidade do analista propor cada hipótese, testá-la, propor uma hipótese substituta ou adicional, testá-la e assim sucessivamente, e desta forma interactiva, criar o modelo.

Esta responsabilidade não desaparece inteiramente com data mining, mas,

- muito do trabalho, encontar o modelo apropriado, é deslocado do analista para o computador.
- O sistema toma a iniciativa da análise de dados, não o utilizador.

Benefícios:

- gerar o modelo requer menor esforço manual (mais eficiente);
- podem avaliar-se muito mais modelos, aumentando assim a possibilidade de encontar melhor modelo;
- o analista necessita de muito menor habilidade, dado que muitos dos procedimentos passo-a-passo são automáticos.

Definição de Data Mining (1)

Já no 1º capítulo, estabelecemos algumas diferenças entre Data Mining e outras ferramentas utilizadas no domínio da extracção de informação de uma base de dados (data query, reporting e OLAP).

O Data Mining, ou mais genericamente o ECBD, pode ser visto segundo diversas perspectivas:

1. Numa perspectiva de negócio, será:
   - O processo de identificação de padrões e relacionamentos escondidos numa base de dados

     Data Mining: Building Competitive Advantage
   - Extracção de informação de negócio útil a partir de grandes bases de dados.

     Data Warehousing, Data Mining and OLAP

Definição de Data Mining (2)

2. Numa perspectiva funcional:
   - **É a procura de informação valiosa em grandes volumes de dados, resultado da cooperação de esforços humanos e de computadores. Os humanos desenham as bases de dados, descrevem problemas e estabelecem objectivos. Os computadores peneiram os dados, procurando padrões que correspondam aos objectivos.**

     Predictive Data Mining: a practical guide, Weiss S.M, and Indurkhya N.

3. Numa perspectiva mais académica:
   - A **extracção implícita**, **não trivial** de **conhecimentos úteis**, **previamente desconhecidos**, dos **dados**.

     Data Mining, Pieter Adrians, Dolf Zantige
   - O **processo não trivial** de identificação de **padrões válidos**, **novos**, **potencialmente úteis** e **compreensíveis** nos **dados**.

     Frawley, Piatetsky e Matheus, 1991

Definição de Data Mining (3)

Analisemos esta última definição:

Padrão - descrição mais simples do que a enumeração de todos os factos.

Processo - O processo de ECBD compreende, em geral, várias fases, envolvendo: (1) Definição do problema, (2) preparação dos dados, (3) procura de padrões, (4) avaliação dos resultados e (5) refinamento iterativo dos resultados.

Não trivial - O processo deve envolver um certo grau de procura de padrões úteis. (Ex. calcular uma remuneração média dos clientes de uma base de dados sobre empréstimos, embora possa ser útil, não poderá ser entendido como extracção).

Definição de Data Mining (4)

Validade - Os padrões extraídos devem, com um **determinado grau de certeza**, ser válidos para novos dados.

Novidade - A novidade pode ser medida com **referência aos dados** (comparação dos valores correntes com valores prévios ou esperados) ou **ao conhecimento** (comparação de uma nova descoberta com as anteriores).

Utilidade potencial - Os padrões de detectados **devem conduzir potencialmente a acções úteis**.

Ex. num exemplo de empréstimos bancários, seria uma medida do aumento de lucros esperados para o banco em resultado da aplicação da regra de decisão decorrente do padrão obtido.

Definição de Data Mining (5)

Compreensibilidade / Sensibilidade - Um dos objectivos da extracção de conhecimento é **tornar os padrões gerados compreensíveis** com vista a possibilitar uma melhor compreensão dos dados.
Como veremos, há técnicas de DM que são inerentemente mais potentes quanto a esta característica (ex. árvores de decisão - transparentes) do que outras (ex. redes neuronais - opacas).

Medida de Interesse - medida do valor de um padrão, combinando:

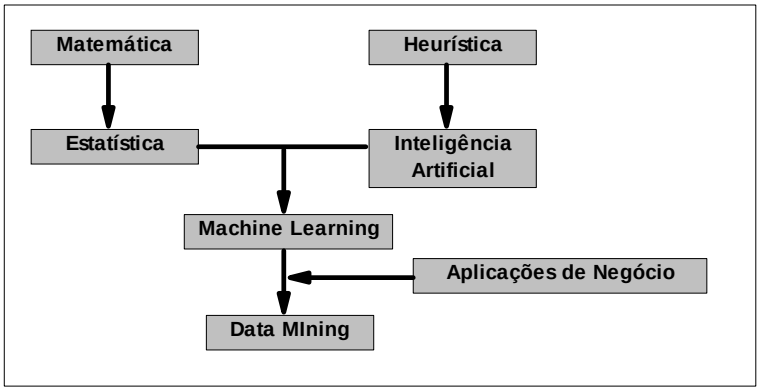
- **validade**
- **novidade**
- **utilidade**
- **simplicidade**

Poder do Data Mining

- O poder do *data mining* é devido ao facto de ele **não depender das vistas humanas estreitas**, para produzir os seus resultados, mas, em seu lugar, **procura e identifica relacionamentos de que os humanos nunca teriam percepção**.
- Uma boa forma de identificar esta realidade é avaliar o modo como um mestre de xadrez distingue um opositor humano de um cibernético.
  - Um computador faz muitas vezes jogadas que um humano nunca executaria, pois que este último "não olhou bem".
  - O que se passa é que a capacidade humana para explorar um grande número de movimentações, num tempo exíguo, é limitada. Tem assim que minimizar a "árvore de pesquisa", limitando o número de caminhos possíveis, baseados na pré-concepção do que entendemos como "estar ou não estar certo".

Breve História do Data Mining

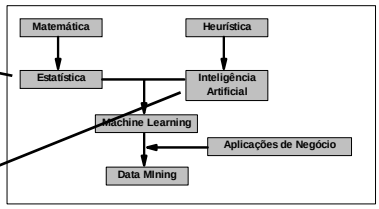
Conceito relativamente recente, o ECBD foi trazido para a ribalta a partir de 1995, aquando da 1.ª conferência internacional sobre KDD, em Montreal. Apesar da sua curta existência, as suas raízes remontam a eras bem mais vetustas (da Estatística e IA).



Breve História do Data Mining

Estatística:

- **Constitui os fundamentos de muitas das tecnologias nas quais o DM é baseado.**
- **Introduz muitos conceitos utilizados para estudar os dados e seus inter-relacionamentos.**



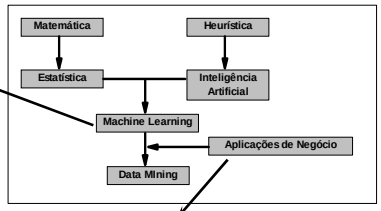
Inteligência Artificial:

- **Tenta aplicar processamento tipo humano (pensamento) aos problemas estatísticos.**
- **Conheceu algumas glórias, mas sofre do chamado problema do "symbol grounding" - mapeamento dos símbolos às entradas sensoriais que é feita pelo intelecto que manipula os símbolos.**
- **Foram criados muitos sistemas periciais bem sucedidos e muitos conceitos de IA foram igualmente adoptados por muitos produtos comerciais no domínios dos SGBDs, nos módulos de optimização de queries.**

Breve História do Data Mining

Machine Learning:

- **Como é mostrado no diagrama, resulta da combinação de heurística da IA com análise estatística avançada.**
- **Tenta fazer com que os programas de computador aprendam sobre os dados que estudam, por forma a que lhes permita tomarem decisões baseadas nas qualidades dos dados estudados, utilizando estatística para os conceitos fundamentais e adicionando heurísticas avançadas e algoritmos para atingir os seus propósitos.**
- **Responde às limitações do processo de aprendizagem de IA - "afunilamento da aquisição de conhecimento", gerando as regras automaticamente a partir da experiência.**
- **Constitui o fundamento de muitas das tecnologias nas quais o DM é construído.**
- **Introduz muitos conceitos utilizados para estudar os dados e seus inter-relacionamentos.**



Machine Learning e o Negócio:

- **Deste encontro surge o Data Mining - trata-se de uma adaptação das técnicas de machine learning a aplicações de negócio.**
- **O DM é melhor descrito como a união de desenvolvimentos históricos e recentes em estatística, AI, e machine learning.**
- **Estas técnicas são utilizadas em conjunto para estudar dados e encontrar padrões escondidos e tendências, previamente desconhecidos.**

Passos na evolução até ao *Data Mining*

Passo de Evolução	Questão típica de negócio	Tecnologias que o permitiram	Empresas que o disponibilizaram	Características
Recolha de Dados (1960s)	"Quais foram os meus lucros totais, nos últimos 5 anos?"	Computadores, tapes, discos	IBM, CDC, ...	Distribuição de dados estática e retrospectiva
Acesso aos Dados (1980s)	"Quais foram as vendas, na zona centro, no mês passado?"	Bases de dados relacionais (RDBMS), SQL, ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Distribuição dinâmica de dados a nível de registo e retrospectiva
Data Warehousing e Suporte à Decisão (1990s)	"Quais foram as minhas vendas, na zona centro, no mês passado? Detalha a seguir só as de Viseu."	Processamento analítico on-line (OLAP), bases de dados multidimensionais, data warehouses	Sybase, Red Bricks, Pilot, Comshare, Arbor, Cognos, Microstrategy, Business Objects	Distribuição dinâmica de dados a níveis múltiplos, ainda retrospectiva
Data Mining (actualidade)	"O que é que poderá acontecer às vendas, em Viseu, no próximo mês? Porquê?"	Algoritmos avançados, computadores multiprocessadores, bases de dados maciças	Pilot, SPSS, IBM, SGI, dataMind, Angloss Software, HNC Software, Information Discovery, Thinking Machines e muitos outros.	Distribuição pró-activa de informação relativa ao futuro

Distinção entre IA e Data Mining

Pode parecer que o Data Mining seja uma parte da IA, mas:

- Os sistemas IA lidam com a codificação do pensamento humano num programa de computador - tentando simular a inteligência.
- Os sistema IA são conduzidos pelo conhecimento humano.

Data Mining, Estatística e Machine Learning, são sistemas:

- Conduzidos por dados - aprendem com exemplos da vida real e não com ideias pré-processadas. Utiliza informação histórica (experiência) para aprender.
- Desta forma:
  - estes sistemas são criados de forma automática e fácil,
  - podem ser actualizados rapidamente,
  - são muito menos onerosos na sua construção,
  - não obrigam à existência de um perito intimamente ligado à criação do sistema.

Distinção entre DM e Estatística

Ex. A regressão é utilizada para criar modelos capazes de predizer o comportamento de clientes, baseados em grandes volumes de dados.

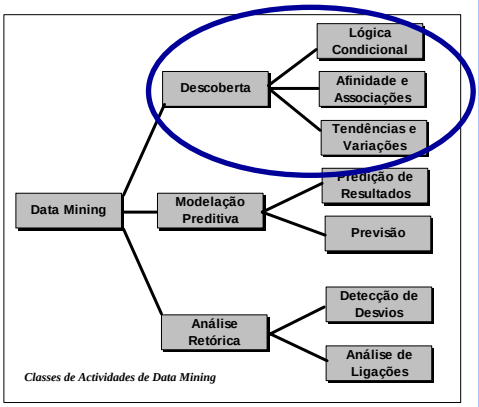
Mas:

- **o DM é capaz de ser utilizado pelo utilizador final**
- **já a estatística, terá de o ser por um perito na matéria**
- **"se a maioria da estatística se traduz num processo de estabelecer uma hipótese e e depois verificá-la, porque não deixar o computador fazer essas tentativas e testá-las automaticamente?"**

Actividades de Data Mining

Descoberta: Processo de procurar na base de dados padrões escondidos sem haver um ideia ou hipótese pré-determinada sobre que padrão será. Encontrar algo que não procuramos.

- Em grandes bases de dados há tantos padrões que o utilizador nunca pensará, na prática, em colocar as questões certas.
- Imagem:
  - temos uma montanha de dados e algures uma pepita de ouro que há que encontar.



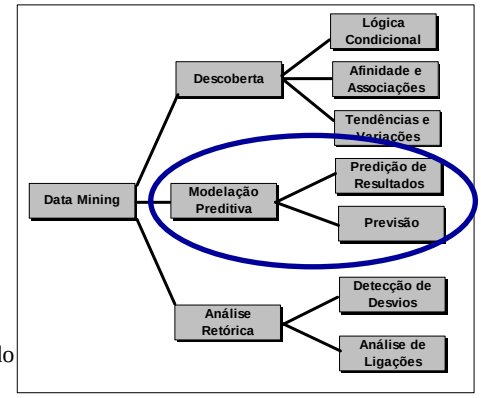
Classes de Actividades de Data Mining

Um cliente que compra guardanapos terá três vezes mais hipóteses de também comprar cerveja

Actividades de Data Mining

Modelação Preditiva: Os padrões descobertos na base de dados são utilizados para predizer o futuro.

- Dado um registo com alguns campos não conhecidos, o sistema tentará adivinhar os valores não conhecidos, baseado nos padrões descobertos previamente na base de dados.
- Dependendo da técnica de DM utilizada, a predição poderá ser "transparente" ou "opaca", dependendo de ser dado ou não ao utilizador o conhecimento da razão para a predição.
- Imagem: conhecer o que conterá a montanha de dados no mês seguinte, ou mesmo a tectónica da montanha...

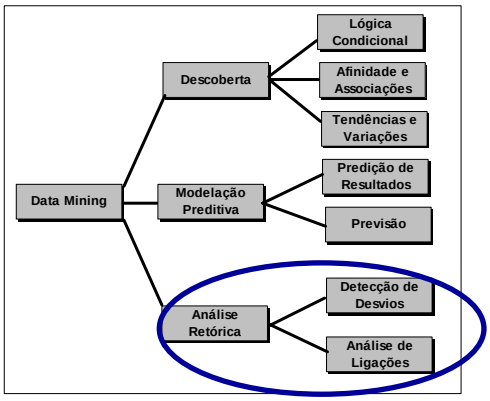


Esta transação é fraudulenta?
Qual o montante do lucro que este cliente gerará?

Actividades de Data Mining

Análise Retórica: É o processo de aplicar os padrões extraídos para encontar elementos de dados anómalos ou não usuais.

Para descobrir o não usual, encontramos primeiro a norma, então detectamos os itens que se desviam do usual dentro de um dado limitar. Os padrões descobertos na base de dados são utilizados para predizer o futuro.



Uma vez descoberto que 97% dos atletas de uma dada base de dados têm menos de 60 anos, poderemos questionar-nos sobre os 3% restantes, saber o porquê (p. ex. praticantes de golfe, onde a idade é menos importante)

Tecnologias de Suporte ao Data Mining

O data mining é possibilitado pela maturidade de quatro tecnologias:

· armazenamento maciço de dados,
· poderosos computadores multiprocessadores,
· algoritmos de datamining,
· visualização de dados.

Fases do Processo de ECBD

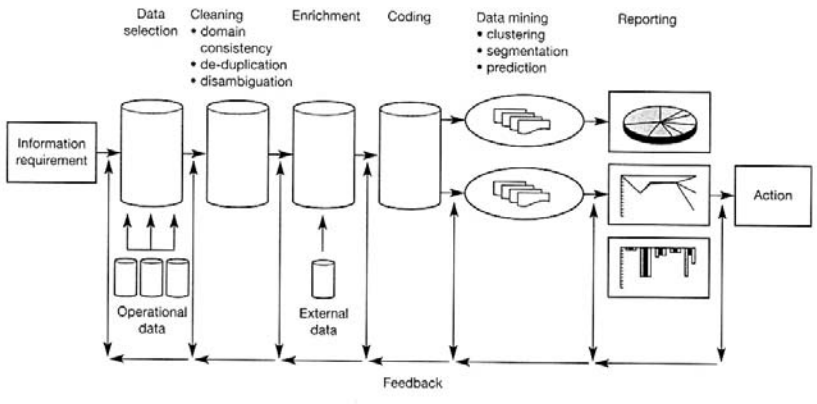
Em princípio, o processo de ECBD consiste em seis estágios:

- Selecção dos Dados
- Depuração
- Enriquecimento
- Codificação
- Extracção de Conhecimento (data mining)
- Relato

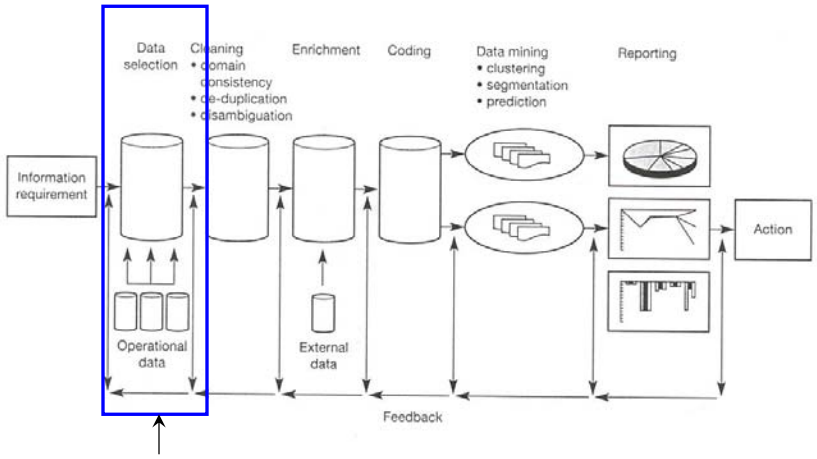
Observações:

- **Embora pareça que há uma trajectória linear, não é o caso. Em qualquer fase, pode ser necessário recuar uma ou mais fases; por exemplo, na fase da codificação ou de extracção de conhecimento, pode suceder apercebermo-nos que a fase de purificação está incompleta, ou descobrir novos dados e utilizá-los para novo enriquecimento.**
- **O processo é contínuo, devendo as organizações trabalhar continuamente os seus dados, identificando constantemente nova necessidade de informação, melhorando os dados para melhor atingir os objectivos.**

Processo de Descoberta



Processo de Descoberta - Selecção

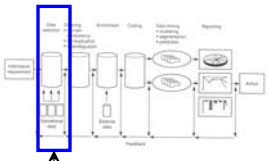


Selecção

Processo de Descoberta - Selecção

Vamos exemplificar as fases do processo de descoberta, utilizando uma pequena base de dados contendo registos dos dados relativos à subscrição de revistas.

Trata-se duma selecção dos dados operacionais do sistema de facturação do editor e contém informação acerca das pessoas que subscreveram revistas.

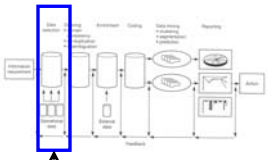


Selecção

Os registos consistem em:

- **n.º de cliente;**
- **nome;**
- **endereço;**
- **data de subscrição;**
- **tipo de revista.**

Processo de Descoberta - Selecção

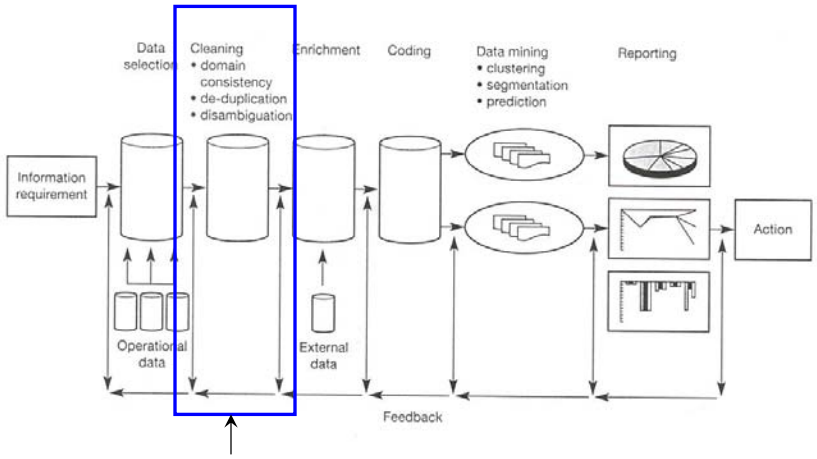


Selecção

Nº de Cliente	Nome	Endereço	Compras Feitas (data)	Revistas Compradas
23003	João	Rua Direita, 3	15-04-94	Automóvel
23003	João	Rua Direita, 3	21-06-93	Música
23003	João	Rua Direita, 3	30-05-92	Humor
23009	Daniel	Rua Formosa, 5	01-01-01	Humor
23013	Pinto	Rua do Arco, 40	30-02-95	Desporto
23019	João	Rua Direita, 3	01-01-01	Casas

Exemplo de registos

Processo de Descoberta - Depuração



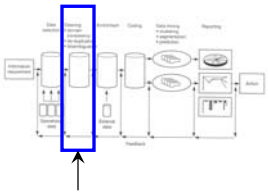
Depuração

Processo de Descoberta - Depuração

É importante para qualquer organização ter consciência da possibilidade de haver anomalias nos seus dados e cuidar de as corrigir.

Apesar do data mining e purificação de dados serem duas disciplinas diferentes, têm muito em comum e utilizando algoritmos de reconhecimento de padrões podemos purificar os dados. Aliás, como vimos numa das fichas das aulas T.Práticas, alguns dos produtos comerciais de purificação de dados, utilizam mesmo este tipo de abordagem.

Há vários tipos de processos de depuração, alguns dos quais podem ser executados em avanço, outros só depois da "poluição" ser detectada, já na fase de codificação ou de extracção.

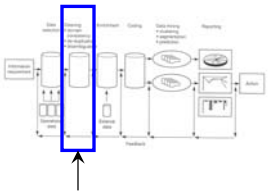


Depuração

Processo de Descoberta - Depuração

Eliminação de duplicações:

- **O senhor João, surge na última linha com o mesmo nome e morada, mas com outro n.º de cliente, pois que o nome está mal escrito.**



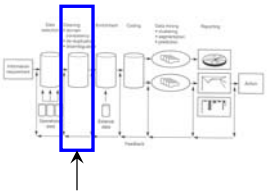
Depuração

Nº de Cliente	Nome	Endereço	Compras Feitas (data)	Revistas Compradas
23003	João	Rua Direita, 3	15-04-94	Automóvel
23003	João	Rua Direita, 3	21-06-93	Música
23003	João	Rua Direita, 3	30-05-92	Humor
23009	Daniel	Rua Formosa, 5	01-01-01	Humor
23013	Pinto	Rua do Arco, 40	30-02-95	Desporto
23019	**Joãoo**	**Rua Direita, 3**	**01-01-01**	**Casas**

Processo de Descoberta - Depuração

Eliminação de inconsistências:

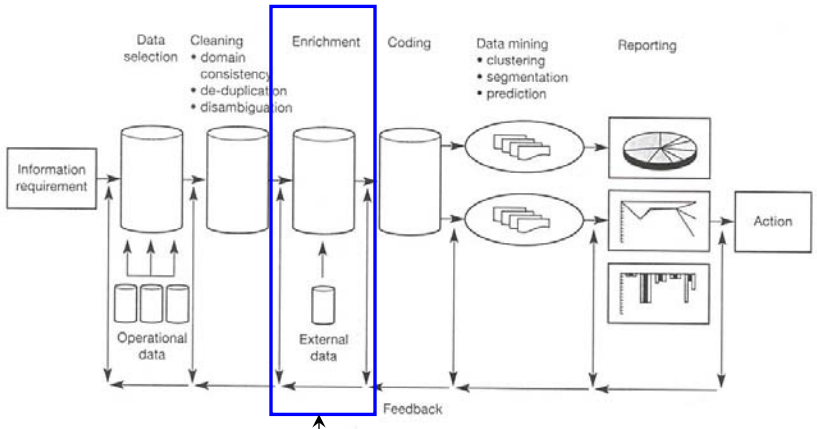
- **Note-se que temos dois registos com data de 1 de Janeiro de 1901, embora a empresa não existisse nessa data. Trata-se claramente de dados incorrectos (outros poderão ter a forma de 11-11-11):**
  - **ou se encontram os dados correctos**
  - **ou se substituem por NULL**



Depuração

Nº de Cliente	Nome	Endereço	Compras Feitas (data)	Revistas Compradas
23003	João	Rua Direita, 3	15-04-94	Automóvel
23003	João	Rua Direita, 3	21-06-93	Música
23003	João	Rua Direita, 3	30-05-92	Humor
23009	Daniel	Rua Formosa, 5	**NULL**	Humor
23013	Pinto	Rua do Arco, 40	30-02-95	Desporto
23019	João	Rua Direita, 3	**30-02-95**	Casas

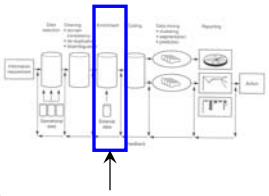
Processo de Descoberta - Enriquecimento



Enriquecimento

Processo de Descoberta - Enriquecimento

Trata-se de tornar os dados mais completos, com mais colunas de valor para o processo de extracção.



Enriquecimento

Processo possível através da:

- **aquisição dados demográficos**
- **realização de entrevistas a subconjuntos de clientes da base de dados, que poderá dar-nos informação detalhada sobre o comportamento dos clientes.**

Neste caso, obtivemos o seguinte:

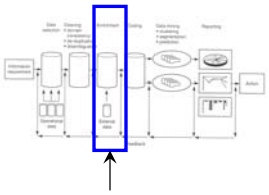
Nome de cliente	Data de nascimento	Rendimento	Crédito	Possui carro	Possui casa
João	13-04-76	$18,500	$17,800	Não	Não
Daniel	20-10-71	$36,000	$26,600	Sim	Não

Obs. De notar que não obtivemos dados relativos ao cliente Pinto.

Processo de Descoberta - Enriquecimento

Adicionando, os dados obtidos, aos registos existentes, obteremos um conjunto de dados com bastante mais conteúdo informativo.

E obteremos a seguinte tabela, depois do enriquecimento:

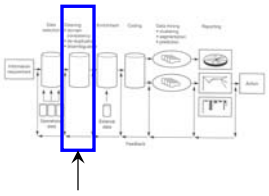


Enriquecimento

Nº de Cliente	Nome	Data de Nasc.	Rendim.	Crédito	Tem Carro	Tem Casa	Endereço	Compras Feitas (data)	Revistas Compradas
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Automóvel
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	21-06-93	Música
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	30-05-92	Humor
23009	Daniel	20-10-71	$26,600	$36,000	Sim	Não	Rua Formosa, 5	**NULL**	Humor
23013	Pinto	NULL	NULL	NULL	NULL	NULL	Rua do Arco, 40	30-02-95	Desporto
23019	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	**30-02-95**	Casas

Processo de Descoberta - Nova Depuração

Antes de dar início à fase seguinte, deveremos seleccionar apenas os registos que contenham suficiente informação ou que tenham valor.



Depuração

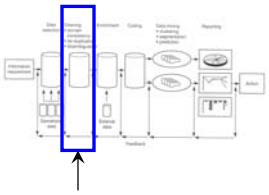
Nº de Cliente	Nome	Data de Nasc.	Rendim.	Crédito	Tem Carro	Tem Casa	Endereço	Compras Feitas (data)	Revistas Compradas
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Automóvel
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	21-06-93	Música
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	30-05-92	Humor
23009	Daniel	20-10-71	$26,600	$36,000	Sim	Não	Rua Formosa, 5	NULL	Humor
23013	Pinto	NULL	NULL	NULL	NULL	NULL	Rua do Arco, 40	30-02-95	Desporto
23019	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	30-02-95	Casas

Processo de Descoberta - Nova Depuração

Antes de dar início à fase seguinte, deveremos seleccionar apenas os registos que contenham suficiente informação ou que tenham valor.

Trata-se duma situação que ocorre frequentemente numa base de dados:

- **muitos dos dados estão em falta**
- **muitos dados são impossíveis de obter.**



Depuração

Que fazer?

- **manter os registos, mas não lhes dar relevância;**
- **removê-los.**

Neste caso, falta-nos informação de especial relevância, relativa ao Snr. Pinto. Desta forma decidimos excluí-lo da amostra final.

N.º de Cliente	Nome	Data de Nasc.	Rendim.	Crédito	Tem Carro?	Tem Casa?	Endereço	Compras Feitas (data)	Revistas Adquiridas
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Automóvel
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	21-06-93	Música
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Humor
23009	Daniel	21-10-71	$26,600	$36,000	Sim	Não	Rua Formosa, 5	NULL	Humor
23013	Pinto	NULL	NULL	NULL	NULL	NULL	Rua do Arco, 40	30-02-95	Desporto
23003	João	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Casas

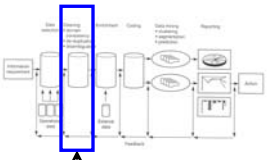
Processo de Descoberta - Nova Depuração

A decisão de remoção é questionável, dado que:

- pode haver uma ligação causal entre a falta de informação e algum comportamento de compra do Sr. Pinto.

Mas vamos supor, por ora, que vamos omitir estes dados sem consequências para o resultado final.

Também o nome do cliente, não deverá ter qualquer influência no seu comportamento de compra, ou seja, trata-se de uma coluna que não terá valor para a solução do problema.

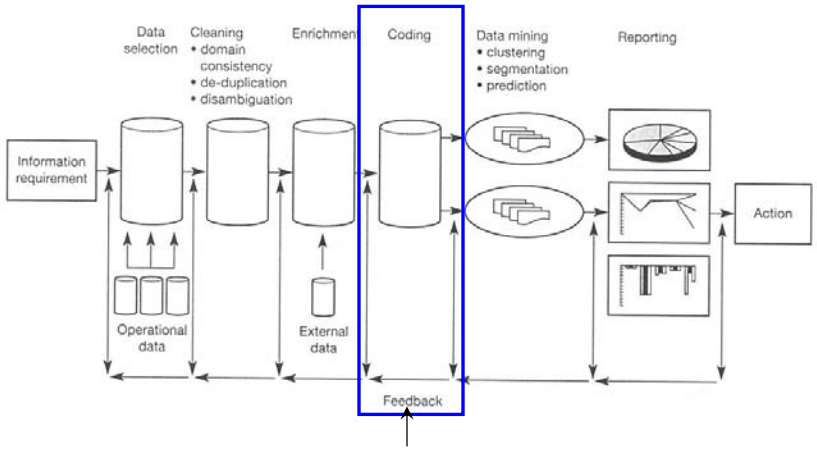


Depuração

Removemos a coluna relativa ao cliente, dado não ter valor para a solução do problema.

N.º de Cliente	Data de Nasc.	Rendim.	Crédito	Tem Carro?	Tem Casa?	Endereço	Compras Feitas (data)	Revistas Adquiridas
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Automóvel
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	21-06-93	Música
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Humor
23009	21-10-71	$26,600	$36,000	Sim	Não	Rua Formosa, 5	NULL	Humor
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Casas

Processo de Descoberta - Codificação



Codificação

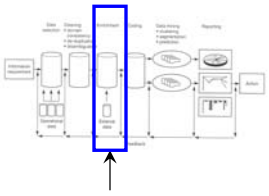
Processo de Descoberta - Codificação

Efectuar transformações criativas nos dados.

Se olharmos a tabela, reparamos que a informação é demasiado detalhada para ser utilizada como entrada para algoritmos de reconhecimento de padrões.

Exemplos:

- **data de nascimento:**
  - **se uma classe for definida para cada data de nascimento, teremos um detalhe demasiado para o nosso propósito!**
  - **mas um algoritmo que opere em classes de idades com um intervalo de, por ex. 10 anos, será muito mais aplicável!**
- **endereços:**
  - **muito detalhe, utilizar códigos de regiões**



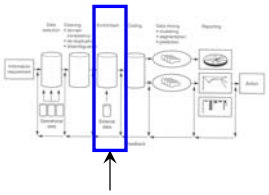
Codificação

A forma como codificamos os dados vai determinar , em grande medida, o tipo de padrões que descobriremos!

Processo de Descoberta - Codificação

A codificação deve ser uma actividade que deve ser repetida sucessivamente, por forma à obtenção dos melhores resultados!

Trata-se dum processo criativo - pode haver um número infinito de códigos diferentes que estejam relacionados com qualquer número de diferentes padrões potenciais que desejaríamos encontrar.



Codificação

Um exemplo:

data de subscrição ⇨ p. ex. convertê-la em número de meses desde 1990 (para encontra padrões sob a forma de séries temporais)

data de subscrição ⇨ p. ex. se interessados na influência sazonal no comportamento do cliente, iremos recodificar as datas de subscrição em códigos sazonais.

Processo de Descoberta - Codificação

Com o 1º exemplo de codificação, relativa à data de subscrição, poderíamos obter a regra:

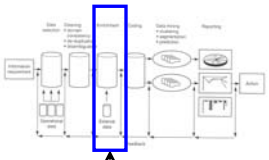
Um cliente com crédito > 13,000 e idade entre 22 e 31 anos que subscrever revista de humor na data T terá grande possibilidade de subscrever revistas de automóveis 5 anos depois.

Ou tendência:

O número de revistas de casas vendidas a clientes com créditos entre 12,000 e 31,000 a viver na região 4 está acrescer.

Ou migração de clientes tipo:

Um cliente com crédito entre 5,000 e 10,000 que lê revistas de humor terá grande possibilidade de tornar-se um cliente com crédito entre 12,000 e 31,000 que lê revistas desportivas e de casas, 12 anos depois.



Codificação

Processo de Descoberta - Codificação

Em resumo, relativamente ao nosso exemplo, podemos aplicar os seguintes passos de codificação:

1. **Endereço para região (em 4 áreas)**
2. **Data de nascimento para idade (100 classes)**
3. **Dividir o rendimento por 1000 (além de simplificar, vai criar classes de rendimentos da mesma ordem de grandeza da idade - 10 a 100)**
4. **Dividir o crédito por 1000**
5. **Converter tem carro e casa de sim/não, para 1/0, facilita a execução de algoritmos**
6. **Converter data de subscrição em meses desde 1990, permitindo executar análises em séries temporais.**

N.º de Cliente	Data de Nasc.	Rendim.	Crédito	Tem Carro?	Tem Casa?	Endereço	Compras Feitas (data)	Revistas Adquiridas
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Automóvel
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	21-06-93	Música
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Humor
23009	21-10-71	$26,600	$36,000	Sim	Não	Rua Formosa, 5	NULL	Humor
23003	13-04-76	$18,500	$17,800	Não	Não	Rua Direita, 3	15-04-94	Casas

Processo de Descoberta - Codificação

Tabela depois de aplicar as transformações indicadas atrás

N.º de Cliente	Idade	Rendim.	Crédito	Tem Carro?	Tem Casa?	Região	Mês de Subscrição	Revistas Adquiridas
23003	20	18.5	17.8	0	0	1	52	Automóvel
23003	20	18.5	17.8	0	0	1	42	Música
23003	20	18.5	17.8	0	0	1	29	Humor
23009	25	26.6	36.0	1	0	1	NULL	Humor
23003	20	18.5	17.8	0	0	1	48	Casas

Neste caso o editor não está interessado em fazer uma análise de séries temporais, ou seja, está mais interessado em relacionamentos entre leitores dos diferentes tipos de revistas do que em relacionamentos entre as revistas e datas de subscrição.

Ou seja, estamos interessados em investigar ligações entre classes de produtos. Então podemos ignorar as datas de subscrição.

Desta forma, a tabela acima, onde por cada subscrição temos uma linha. É de difícil análise: Mas pode transformar-se, por aglutinação de linhas correspondentes ao mesmo cliente, mostrando as revistas subscritas ou não por ele.

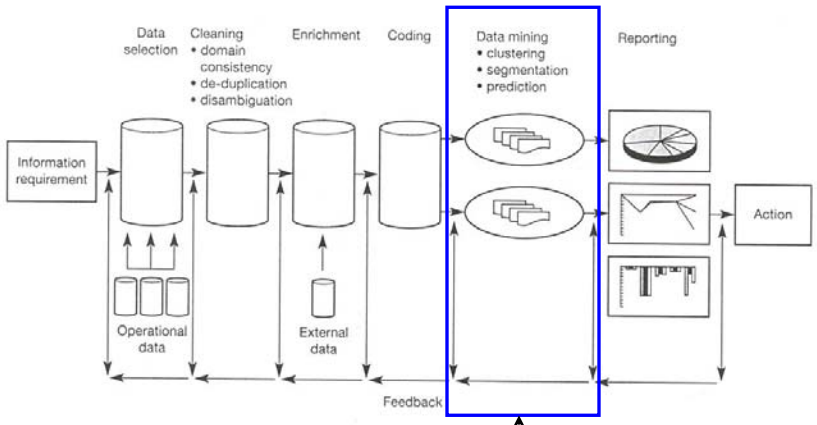
Processo de Descoberta - Codificação

N.º de Cliente	Idade	Rendim.	Crédito	Tem Carro?	Tem Casa?	Região	Mês de Subscrição	Revistas Adquiridas
23003	20	18.5	17.8	0	0	1	52	Automóvel
23003	20	18.5	17.8	0	0	1	42	Música
23003	20	18.5	17.8	0	0	1	29	Humor
23009	25	26.6	36.0	1	0	1	NULL	Humor
23003	20	18.5	17.8	0	0	1	48	Casas

N.º de Cliente	Idade	Rendim.	Crédito	Tem Carro?	Tem Casa?	Região	Auto	Casas	Desp.	Música	Humor
23003	20	18.5	17.8	0	0	1	1	1	0	1	1
23009	25	26.6	36.0	1	0	1	0	0	0	0	1

A última operação realizada sobre o campo revistas subscritas, é denominada de flattening - um atributo com cardinalidade n é substituído por n atributos binários

Processo de Descoberta - Extracção



Extracção

Processo de Descoberta - Extracção

A fase de extracção do processo de ECBD (data mining) emprega um conjunto de técnicas. Embora várias técnicas diferentes sejam utilizadas para própósitos diferentes, as que são de interesse no contexto presente são:

- **Interrogação da base de dados**
- **Análise estatística**
- **Visualização**
- **Processamento analítico em linha (OLAP)**
- **Aprendizagem baseada em casos (K-nearest neighbor)**
- **Árvores de decisão**
- **Regras de associação**
- **Redes neuronais**
- **Algoritmos genéticos**

Extracção - Análise preliminar dos dados

Recurso a uma linguagem de interrogação (query tool)

- **Trata-se de efectuar uma análise grosseira do conjunto de dados**
- **Através da utilização de SQL poderemos obter informação valiosa relativamente ao data set**
  - **já foi dito atrás que cerca de 80% da informação valiosa pode retirar-se através de query**
  - **só os restantes 20% obrigam à utilização de técnicas mais avançadas.**

Para os exemplos seguintes, vamos utilizar uma base de dados relativa à subscrição de revistas (1000 clientes), de onde foi retirado o set até aqui utilizado.

Extracção - Análise preliminar dos dados

Os valores médios, apresentam-se na tabela abaixo:

	Média
Idade	46.9
Rendimento	20.8
Crédito	34.9
Carro próprio	0.59
Casa própria	0.59
revista de carros	0.329
revista de casas	0.702
revistas de desporto	0.447
revista de música	0.146
revista de humor	0.081

Os números relativos à média são muito importantes, pois que nos dão uma norma que permite ajuizar do desempenho dos algoritmos de reconhecimento de padrões.

Um algoritmo que prediga sempre a <u>não subscrição</u> de revistas de automóveis, estará correcto em 671 por 1000 casos, cerca de 70%.

Qualquer algoritmo que reclame obter alguma visão sobre estes dados, permitindo alguma predição real, deverá melhorar este valor.

Predição Naïve - Resultado trivial que é obtido através dum método extremamente simples. O algoritmo de aprendizagem deve fazer melhor do que isto.

Extracção - Análise preliminar dos dados

Predições Naïve

Revistas	probabilidade *a priori* que os clientes subscrevam a revista	Precisão de predição Naïve
revista de carros	32.9%	67.1%
revista de casas	70.2%	70.2%
revistas de desporto	44.7%	55.3%
revista de música	14.6%	85.4%
revista de humor	8.1%	91.9%

É mais difícil fazer predições relativamente ao grupo mais pequeno do nosso conjunto de dados.

Um algoritmo que prediga que um cliente subscreverá revistas de humor, terá de efectuar predições com precisão superior a 92% (precisão da predição Naïve).

Extracção - Análise preliminar dos dados

Resultados da aplicação de predição Naïve (Médias)

Revista	Idade	Rendim.	Crédito	Carro	Casa
carros	29.3	17.1	27.3	0.48	0.53
casas	48.1	21.1	35.5	0.58	0.76
desporto	42.2	24.3	31.4	0.70	0.60
música	24.6	12.8	24.6	0.30	0.45
humor	21.4	25.5	26.3	0.62	0.60

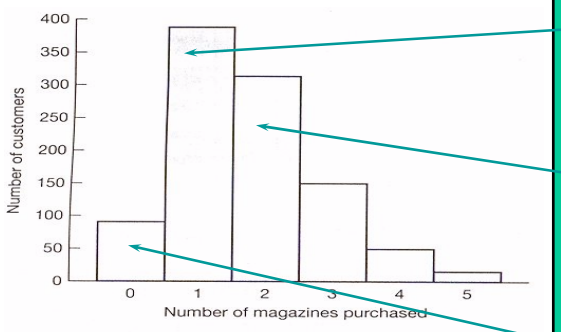
Análise breve:

a média de idade de leitores de revistas de carros é inferior à média de idades dos clientes

a média de idade dos leitores de revistas de humor é muito baixa

Extracção - Análise preliminar dos dados

Número de subscritores múltiplos de revistas



Análise breve:

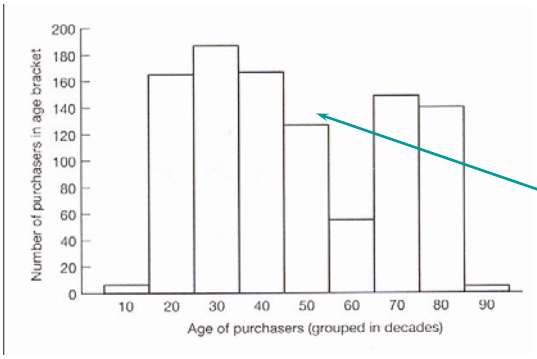
quase 40% dos clientes compram pelo menos uma revista

31% subscrevem 2 revistas (podendo indicar que pode haver padrões interessantes a descobrir entre grupos de subscritores simples e múltiplos)

há cerca de 9% de clientes que não o são efectivamente (resultado de possíveis erros na base de dados?)

Extracção - Análise preliminar dos dados

Idade dos subscritores (agrupados em décadas)



Análise breve:

uma distribuição quase igual (fora os muito novos e muito velhos)

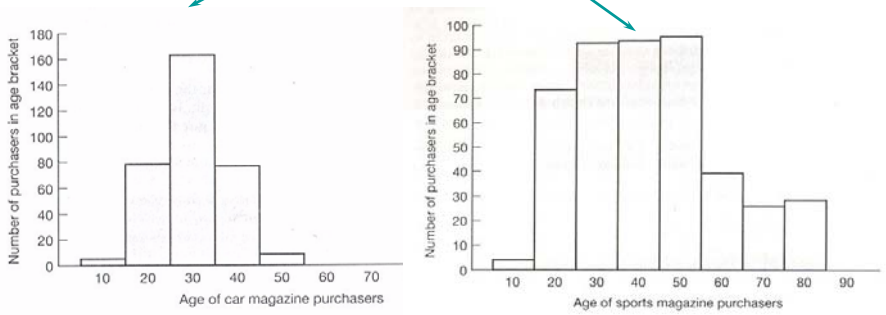
Extracção - Análise preliminar dos dados

Idade dos subscritores, avaliada por subgrupos (neste caso subscritores de revistas de carros e de desporto)

Análise breve:

subscritores de revistas de carros com idades à volta dos 30 anos

leitores de revistas de desporto com distribuição mais alargada



Extracção - Visualização

As técnicas de visualização constituem um método muito útil de descoberta de padrões.

Podem ser utilizadas no início do processo de data mining:

- **obter um feeling da qualidade dos dados**
- **onde poderão ser encontrados os padrões.**

Algumas ferramentas mais sofisticadas permitem:

- **a exploração de estruturas tri-dimensionais interactivamente**
- **navegar através de espaços artificiais de dados**
- **a evolução dos dados podem ser mostrados sob a forma de filme animado**

No entanto, muitas das vezes teremos de nos limitar a técnicas muito mais simples, que, de qualquer forma, nos fornecem informação valiosa.

- **Exemplo: diagrama de dispersão**

Extracção - Visualização

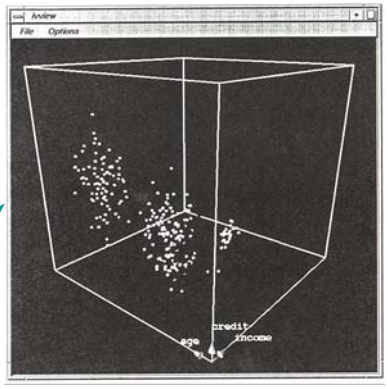
Rendimento X Idade dos subscritores de revistas de música

Análise breve:

pessoas novas e com baixos rendimentos tendem a ler revistas de música

Extracção - Visualização

- **A metáfora espacial é muito útil em data mining.**
- **Permite-nos determinar a distância entre dois registo no espaço de dados:**
  - **registos que estejam muito próximos são muito semelhantes;**
  - **registos muito distantes, representarão indivíduos que terão muito pouco em comum.**



Outra vantagem da boa codificação: por forma a obter uma boa comparação entre valores, deveremos normalizar os atributos.

Ex. se idade 1-100 e rendimento 0-100,000, este último seria um atributo muito mais distintivo do que a idade, não sendo o pretendido (as escalas deverão ter a mesma ordem de valores).

Análise de distâncias entre registos no espaço de dados, relativo aos atributos: idade, rendimento e crédito

Extracção - Visualização

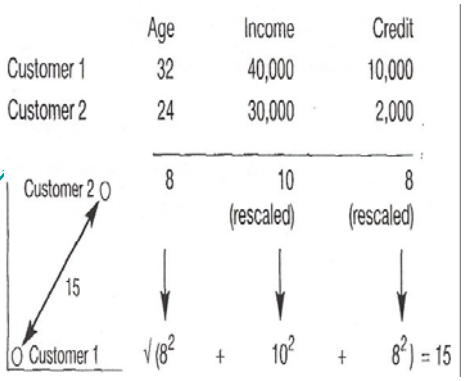
- **Os registos formam pontos no espaço determinados pelos seus atributos, e a distância entre eles pode ser medida.**
- **Suponhamos dois clientes:**
  - **Cliente 1, idade 32, rendimento 40,000 e Crédito 10,000**
  - **Cliente 2, idade 24, rendimento de 30,000 e crédito 2,000.**

	Age	Income	Credit
Customer 1	32	40,000	10,000
Customer 2	24	30,000	2,000

Utilizando uma medida de distância euclidiana, a distância entre o cliente 1 e 2 será de 15 (com normalização).

Idade= 32-24=8, Rendimento (40-30=10), Crédito (10-2)=8

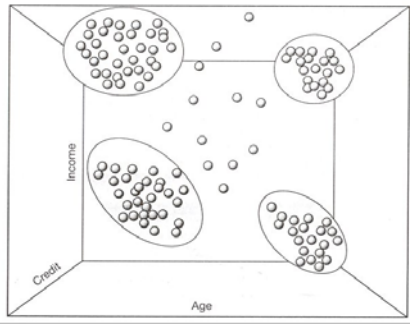
Distância Euclidiana=$\sqrt{8^2+10^2+8^2}$=15



Extracção - Visualização

- **Se tornarmos o intervalo de valores de todas as medidas idêntico, obteremos uma medida de distância fiável entre os diversos registos.**
- **Predições interessantes podem ser visualizadas utilizando programas de pesquisa avançados, identificando visualmente clusters de clientes potenciais que terão grande possibilidade de adquirir um certo produto.**

Clusters interessantes relativamente ao nosso exemplo.



•Para espaços de dados de baixa dimensionalidade é fácil a identificação de nuvens de dados, e muitas vezes, identificar clusters interessantes através de mera inspecção visual.

Extracção - Ferramentas OLAP

- **Como já vimos no capítulo 2, estas ferramentas permitem o acesso a diversas formas de análise e visualização multidimensional e interactiva da informação.**
- **Aproximam a análise à forma de ver o negócio - um determinado valor é visto segundo um conjunto de perspectivas lhe que fornecem caracterização.**
- **Funcionalidades dum sistema OLAP:**
  - **cálculo e modelação multidimensional**
  - **análise de tendências em sequências temporais**
  - **análise de agrupamento de dados**
  - **movimentação e comparações ao longo das dimensões em consideração**
- **<u>Observação Importante</u>: as ferramentas OLAP não aprendem, não criam qualquer novo conhecimento, não pesquisam novas soluções. Obrigam, em regra, a motores multidimensionais intermédios ou a novas formas de armazenamento dos dados.**

Aplicar Data Mining (1)

Classificação e Regressão

- **Representam a maioria dos problemas a que são aplicadas as técnicas de data mining, criando modelos capazes de predizer o valor ou classe de uma variável dependente, utilizando técnicas de indução supervisionada.**
- **São criados modelos :**
  - **de classificação - se capazes de predizer a classe a que um determinado membro pertence;**
  - **de regressão - se capazes de predizer um valor.**
- Em classificação, a resposta é simplesmente verdade ou falso; já em regressão, a resposta é um número, como, por exemplo, os lucros ou perdas relativos a um empréstimo.

"O objectivo desta operação é utilizar o conteúdo da base de dados (dados sobre o passado) para gerar automaticamente um modelo que possa predizer o comportamento futuro."

"IBM's Data Mining Technology"

Aplicar Data Mining (2)

Classificação e Regressão (continuação)

- **Exemplos**
  - **predizer se um determinado empréstimo será ou não um bom risco de crédito;**
  - **predizer a rentabilidade de um cliente**
  - **predizer a probabilidade de que um determinado paciente tenha uma dada doença**
- As séries temporais são apenas um tipo especial de problema relativo a regressão ou classificação, onde as medidas são obtidas a intervalos de tempo, como será o caso dos pagamentos relativos a um empréstimo.

Aplicar Data Mining (3)

Associação e Sequência (também chamados de análise de cesto de compras)

- **O objectivo é o estabelecimento de relações entre os registos de uma base de dados, não a criação de um modelo que caracterize o conteúdo de uma base de dados, como no processo anterior.**
- **Geram modelos descritivos que descobrem regras como:**
  - **os clientes que compram espaguete têm três vezes mais possibilidade de adquirirem queijo do que aqueles que não compram.**
    - Exemplo, um gestor de vendas de um supermercado está muito interessado em conhecer que produtos se vendem conjuntamente, por forma a adquiri-los para a loja e a dispô-los fisicamente perto, implicando porventura um reordenamento das respectivas secções, mas decerto visando um incremento das vendas, pois que o cliente será automaticamente lembrado da conveniência da compra adicional.
- Trata-se duma operação que é suportada por técnicas de descoberta de associações e sequências.
  - A resposta deste tipo de operação aos problemas que lhe são colocados tem um formato lógico como "Um empréstimo é pago com 90% de confiança, quando o proprietário tem um história pessoal de pagamento atempado dos débitos efectuados com cartão de crédito".

Aplicar Data Mining (4)

Clustering (segmentação da base de dados)

- **Técnica descritiva que agrupa entidades similares em conjunto, colocando entidades dissimilares em grupos diferentes.**
- **Pode ser utilizada em marketing para encontrar grupos de clientes com afinidades e, em cuidados de saúde, para encontar pacientes com perfis semelhantes.**
- **Esta operação surge como resultado de necessidade de obter-se um resumo de cada base de dados ou antes de dar início a uma das operações de *data mining*.**
- **O Clustering é muito subjectivo:**
  - **Dado que se emprega uma medida de distância, como a técnica de vizinho mais próximo (nearest neighbor), os clusters estão completamente dependentes da medida da distância que é utilizada.**
  - **Normalmente é de interesse o envolvimento de um perito no domínio de análise pretendido, para propor a medida de distância apropriada.**

Aplicar Data Mining (5)

Clustering (segmentação da base de dados)

Um exemplo de utilização será o de uma cadeia de armazéns que mantém o registo das compras efectuadas pelos seus clientes, conhecendo as compras efectuadas, em qualquer visita de um seu cliente a cada uma das lojas. Neste caso, será muito útil segmentar a base de dados, baseando a divisão em períodos significativos de análise, como: período de "regresso às aulas", "antes do Natal", etc. Sobre cada um destes segmentos, poderá depois ser aplicada análise de ligações, para identificar que produtos são vendidos em conjunto. Para executar esta operação, são utilizadas técnicas de agrupamento "*clustering*".

Aplicar Data Mining (6)

Detecção de Desvios

- Trata-se de uma operação no quadrante oposto da anterior, mas relacionada com ela.
  - Atrás, era importante identificar registos relacionados e estabelecer, com isso, grupos e as correspondentes divisões;
  - Agora, normalmente depois da segmentação efectuada, há que "identificar pontos que caem fora de um conjunto de dados particulares, explicando se serão devidos a ruído ou outras impurezas, ou por razões casuais";
  - É especialmente devido a estas últimas que esta operação é importante: em muitos casos, a identificação e explicação de um desvio é a fonte de descobertas puras, pois que expressam desvios em relação a expectativas e normas conhecidas previamente, podendo indicar o surgir de novas tendências ou oportunidades. Esta operação é suportada por técnicas estatísticas, tais como o teste de significância, onde sumarizações de estatísticas (média e desvio standard) são utilizadas para medir as diferenças.

Aplicar Data Mining (7)

- **Text Mining**.
  - Também conhecido por Text Data Mining ou Descoberta de Conhecimento em Bases de Dados Textuais
  - Refere-se ao processo de extracção de padrões de interesse e não triviais ou conhecimento a partir de documentos de texto não estruturados.
  - Pode ser visto como uma extensão ao Data Mining ou KDD.
  - Dada a abundância de documentos (é a mais normal forma de arquivar informação – um estudo recente indica que cerca de 80% da informação das empresas reside sob a forma de documentos) possui um enorme potencial (maior mesmo do que o DMining).

Aplicar Data Mining (8)

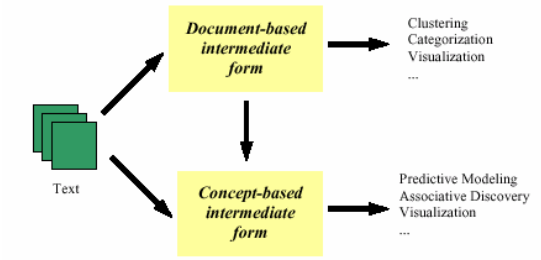
- **Text Mining**.
  - Trata-se duma operação de aplicação de métodos de *text-mining* a informação textual (campos texto existentes em registos, até documentos inteiros).
  - Difícil pois que a informação textual não é vocacionada a ser manipulada por computadores. Os documentos têm uma estrutura interna limitada, se a tiverem. Os dados sob a forma de texto são inerentemente não estruturados e difusos. São ricos semanticamente, se tratados como um todo. Além disso, a informação importante que contêm não é explícita, mas implícita, dispersa por todo o texto.
  - O texto terá de ser transformado para um formato adequado à aplicação posterior dos métodos respectivos.

Aplicar Data Mining (9)

- **Text Mining**.
  - O Text Mining é assim uma campo multidisciplinar, envolvendo:
    - Recuperação de informação
    - Análise de textual
    - Clustering
    - Extracção de informação
    - Categorização
    - Visualização
    - Tecnologias de bases de dados
    - Machine learning
    - Data Mining

Aplicar Data Mining (10)

- **Referêncial p/ Text Mining (2 fases):**
  - **Refinamento de texto**
  - **Destilação de conhecimento**



Converte texto não estruturado (documentos de texto em forma livre) numa forma intermédia (IF)

A IF pode ser:

- *Baseada em documento, em que cada entidade reprsenta um documento*
- *Baseada em conceitos, em que cada entidade representa um objecto ou conceito de interesse para um domínio específico.*

Text refining

Knowledge distillation

Deduz padrões ou conhecimento a partir da forma intermédia.

- *A partir de IF baseada em documento deduz padrões e relacionamentos entre documentos*
- *A partir de IF baseada em conceitos deriva padrões e relacionamentos entre objectos ou conceitos*

Retirado de Text:Mining – The State of the art and the challenges – Ah-Hwee Tan

Aplicar Data Mining (11)

Text Mining (continuação).

O *Text Mining* envolve, pois, dois aspectos:

- **pesquisa de informação que permite aos utilizadores encontrar a informação de que necessitam;**
- **ferramentas de análise de texto que ajudam a extrair conhecimentos chave do texto, organizar documentos por assunto, e encontrar temas predominantes num conjunto de documentos.**

A utilização prática do primeiro destes aspectos permite a construção de sistemas de alta qualidade de pesquisa de dados, com aplicação alargada, desde a pesquisa de documentos da empresa, quer na Web.

Quanto ao segundo aspecto, análise de texto, é executado por um conjunto de operações sobre o texto mais semelhantes ao *data mining*. Permitem identificação da linguagem, reconhecimento de itens de vocabulário significativos num texto, geração de agrupamentos de documentos, categorização que atribui documentos a categorias pré-existentes.

Aplicar Data Mining (12)

Produtos de Text Mining Ilustrativos

Company/ Organization	Product/ Application	Text Refining Functions	Intermediate Form	Knowledge Distillation Functions
Cartia	ThemeScape		Document-based	Clustering, visualization
Canis	cMap		Document-based word histograms	Clustering, visualization
IBM/ Synthema	Technology Watch		Document-based	Clustering, visualization
Inxight	VisControls		Document-based Hyperbolic tree	Visualization
Semio Corp	SemioMap		Concept-based	Visualization
Knowledge Discovery System	Concept Explorer	Info retrieval	Concept-based	
Inxight	Linguist	Info retrieval, text analysis, summarization	Document-based	
IBM	iMiner	Info retrieval, summarization	Document-based	Clustering, categorization
TextWise	DR_LINK CINDOR CHESS	Info retrieval Info extraction	Concept-based	
Cambio	Data Junction	Info extraction	Concept-based	
Megaputer	TextAnalyst	Info retrieval, summarization	Document-based semantic net	Classification

Retirado de Text:Mining – The State of the art and the challenges – Ah-Hwee Tan