



SIMILARIDADE DE DADOS

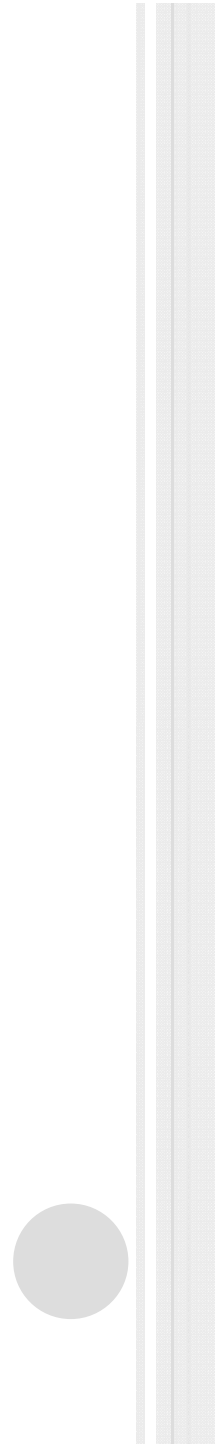
Carina F. Dorneles
dorneles@inf.ufsc.br

SIMILARIDADE – O QUE É?

- Mostrar computacionalmente, através de um **valor**, o quanto dois objetos são semelhantes entre si
- Entende-se por “**Objeto**”
 - Texto simples (uma palavra)
 - Tuplas (linhas) de uma tabela de um BD
 - Linhas de uma tabela Web
 - Textos longo
 - Documentos estruturados (XML, HTML)
 - Documentos não-estruturados (txt, códigos de programa, blogs)
 - Estruturas de dados mais complexas
 - Árvores
 - Grafos
 - Tuplas
 - Imagens, sons, vídeos



SIMILARIDADE

- Conceito subjetivo
 - O que é similar para **algumas pessoas** pode não ser para **outras**
 - As métricas de cálculo de similaridade são bastante distintas
 - Usam diferentes mecanismos para gerar o grau de similaridade
- 

CALCULANDO A SIMILARIDADE

- Exemplo:
 - Para textos curtos (atributos de tabelas – BD ou Web):
 - Levenshtein
 - Transformação - inserção, deleção e substituição
 - LCS
 - sequência comum mais longa
 - Jaro-Winckler
 - Fórmula que envolve transposição de n sequências comuns encontradas- Cada uma gera um score, que tem distribuições diferentes sobre um mesmo domínio de valores

CALCULANDO A SIMILARIDADE

- Exemplo:
 - Para textos longos (documentos semi-estruturados ou não-estruturados):
 - Diff
 - Diferenças entre dois documentos
 - XDiff
 - Detecção de diferença em árvores
- Cada uma gera um score, que tem distribuições diferentes sobre um mesmo domínio de valores

CALCULANDO A SIMILARIDADE

- A ideia básica:
 - Função **recebe um par de objetos**
 - FuncaoDeSimilaridade (parametro1, parametro2)
 - Função **retorna um escore** que indica quão similares são os parâmetros

DIFERENTES ÁREAS DIFERENTES FOCOS

- Aplicações com dados textuais e/ou estruturados
 - Fazem desambiguação
 - Identificar o que trata o **mesmo objeto do mundo real**
 - Usa as funções de similaridade como um substituto do operador de igualdade
- Aplicações com dados binários (imagens, sons, vídeos)
 - Consultar **objetos similares**
 - Não necessariamente o mesmo
- Aplicações de tomada de decisão
 - Identificar **objetos similares**
 - Que objeto *se comporta* de forma similar ao outro

DIFERENTES ÁREAS DIFERENTES FOCOS

- Aplicações com dados textuais e/ou estruturados
 - Fazem desambiguação
 - Identificar o que trata o **mesmo**
 - Usa as funções de similaridade/igualdade
- Aplicações com dados estruturados
 - Consultar **objetos similares** (objetos)
 - Não necessariamente o mesmo
- Aplicações de tomada de decisão
 - Identificar **objetos similares**
 - Que objeto *se comporta* de forma similar ao outro

Inúmeras propostas de
funções/abordagens/
algoritmos de similaridade

EXEMPLO COM DADOS FEEDBACKS

```
<artigo>  
  <titulo>Avanços TI </titulo>  
  <secao>De acordo com a  
(ONU), os avanços tecnológicos nos  
países sub-desenvolvidos não  
cresceu como deveria...  
  <secao>  
</artigo>
```

```
<artigo data='2008/01/01'>  
  <titulo>Avanços TI </titulo>  
  <secao>De acordo com a  
Organização das Nações Unidas  
(ONU), os avanços tecnológicos  
nos países sub-desenvolvidos não  
cresceu como deveria...  
  <secao>  
</artigo>
```

EXEMPLO COM DADOS RELEVANTES

Members

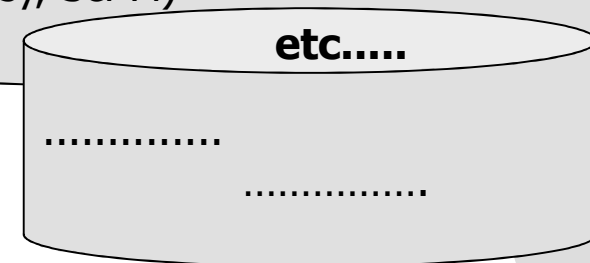
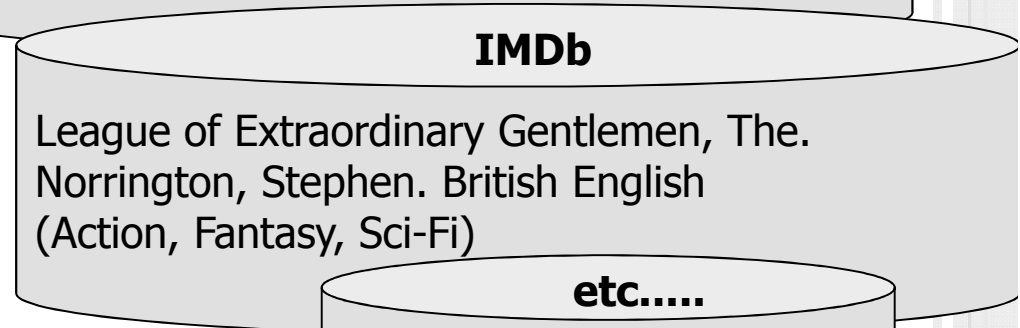
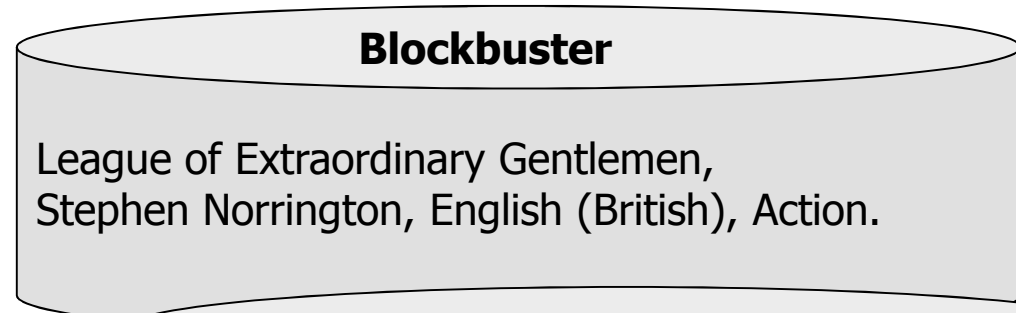
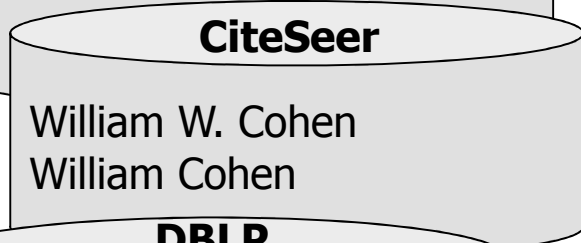
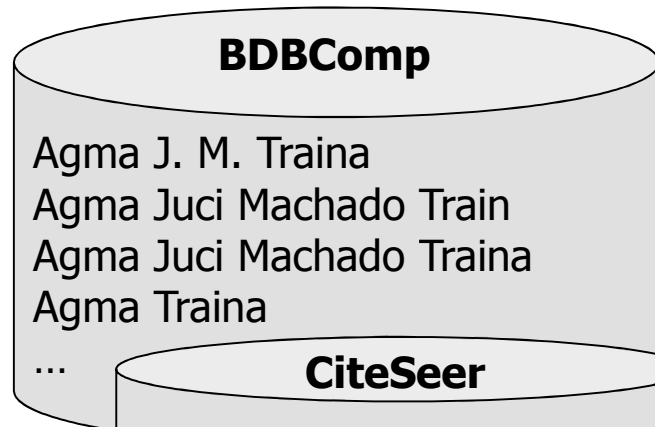
<i>Name</i>	<i>Inst. Repr.</i>	<i>email</i>
Kofi Annan	ONU	kofiannan@...

Members

<i>Name</i>	<i>Institution</i>	<i>e-mail@</i>
Annan, Kofi	United Nations (UN)	kofi@...

EXEMPLO COM DADOS TEXTUAIS

○ Instâncias reais



CONSULTAS

- Como efetuar consulta sobre uma base que possui diferentes representações do mesmo objeto?
 - Exemplo usando um dialeto similar a SQL



```
SELECT artigo  
FROM BDBComp  
WHERE levenshtein(autor, 'Agma Machado Traina') > 0,75
```

Recuperar artigos do
autores 'Agma Machado
Traina'

BDBComp

Agma J. M. Traina
Agma Juci Machado Train
Agma Juci Machado Traina
Agma Traina
... ..

INTEGRAÇÃO DE DADOS

- Como integrar dados que são escritos de diferentes formas?

Levenshtein ("Blockbuster.movie.title", "IMDb.movie.title") ≥ 0.78

Blockbuster

League of Extraordinary Gentlemen,
Stephen Norrington, English (British), **Action.**

IMDb

League of Extraordinary Gentlemen, The.
Norrington, Stephen. British English
(Action, Fantasy, Sci-Fi)

EXEMPLO COM IMAGENS

○ Exemplo

Consulta Exemplo



- Cada objeto (imagem) é representado através de características
- Características são extraídas
 - Cor
 - Textura
 - Formas geométricas
 - Etc...



EXEMPLO USANDO *CROSS-MODAL SEARCH ENGINE*

<http://dolphin.unige.ch/cmse/>

Consulta Exemplo



EXEMPLO USANDO *CROSS-MODAL SEARCH ENGINE*

<http://dolphin.unige.ch/cmse/>

Consulta Exemplo



Conjunto-resposta



1/514.jpg - [Annotation](#)



7/63419.jpg - [Annotation](#)



24/234537.jpg - [Annotation](#)



2/16204.jpg - [Annotation](#)



12/116503.jpg - [Annotation](#)



15/142069.jpg - [Annotation](#)



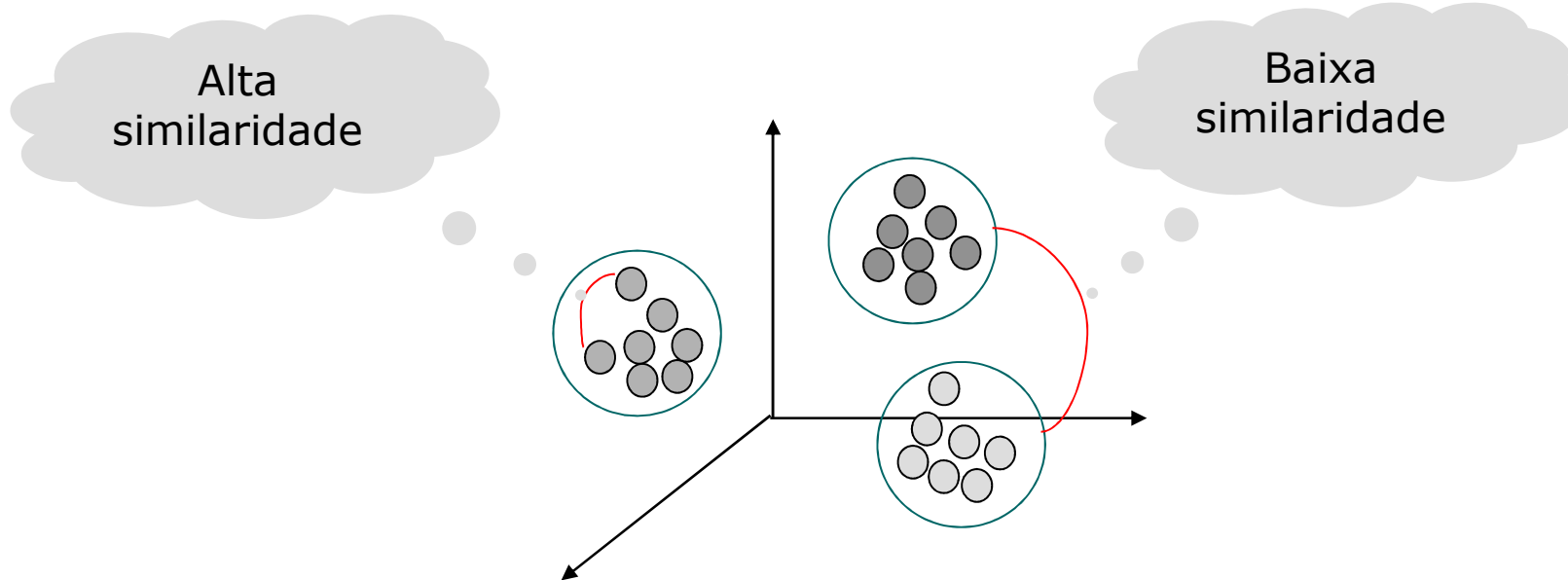
16/153557.jpg - [Annotation](#)



6/55301.jpg - [Annotation](#)

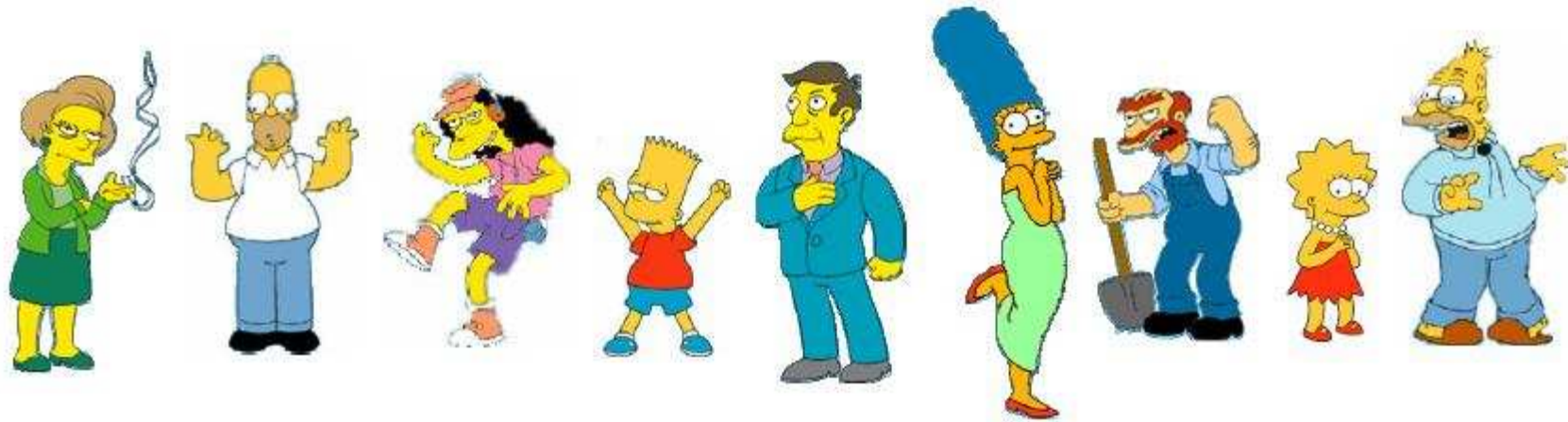
EXEMPLO DE APLICAÇÕES DE DATA-MINING

- Agrupar linhas das tabelas de acordo com a similaridade existente entre elas



EXEMPLO DE AGRUPAMENTO DE DADOS

- Encontrar grupos de clientes similares



MATRIZ DE SIMILARIDADE



1	0.2	0.2	0.3	0.3
	1	0.8	0.6	0.6
		1	0.7	0.7
			1	0.9
				1

$$1 - D(\text{Marge Simpson}, \text{Marge Simpson}) = 1$$

$$1 - D(\text{Marge Simpson}, \text{Bart Simpson}) = 0.8$$

$$1 - D(\text{Maggie Simpson}, \text{Lisa Simpson}) = 0.2$$



Como calcular o grau de similaridade?

SIMILARIDADE VS. DISTÂNCIA

- Uma função de similaridade $fs(a1, a2) \rightarrow s$
 - Escore s no intervalo $[0, 1]$.
 - Quanto **maior** o valor do escore, **mais similares** os dois valores $a1$ e $a2$ são entre si.
- Uma função de distância $fd(a1, a2) \rightarrow s$
 - Escore s no intervalo $[0, \infty]$.
 - Quanto **menor** o valor do escore, **menos similares** os dois valores $a1$ e $a2$ são entre si.

Grande parte das propostas trabalham com distância



SIMILARIDADE TEXTUAL


Valores atômicos





LEVENSHTEIN

- Função Levenshtein()

- Originalmente, é uma **função de distância** que calcula o número de operações necessárias para transformar uma *string* em outra
 - Para usá-la como função de similaridade precisamos
 - Normalizar o valor da distância
 - Reduzir o valor de distância resultante de 1
- 

LEVENSHTEIN

- SimLev = $1 - \frac{\text{Levenshtein}(s1, s2)}{\max(\text{size}(s1), \text{size}(s2))}$
- Exemplo: *s1* = *deterministico* e *s2* = *determinado*

Levenshtein(s1,s2) vai realizar
operações para transformar

**Deterministico
em
Determinado**

LEVENSHTEIN

- SimLev = $1 - \frac{\text{Levenshtein}(s1, s2)}{\max(\text{size}(s1), \text{size}(s2))}$
 - Exemplo: $s1 = \text{deterministico}$ e $s2 = \text{determinado}$

S1 = D E T E R M I N I S T I C O
S2 = D E T E R M I N A D O

D E T E R M I N A D O

Operações:

↓ replace
↓ delete

LEVENSHTEIN

- Assim, temos

S1 = D E T E R M I N I S T I C O
S2 = D E T E R M I N A D O

D E T E R M I N A D O

Diagram illustrating the Levenshtein distance between S1 = DETERMINISTICO and S2 = DETERMINADO. The alignment shows 3 replacements (I to A, S to A, T to O) and 3 deletions (I, C, O).

Operações:

↓ replace
↓ delete

- 6 operações (3 *replaces* e 3 *deletes*)
- **Levenshtein (s1, s2) = 6**
- Para transformar o valor em similaridade
 - $\max(\text{size}(s1), \text{size}(s2)) = 14$
 - Similaridade = $1 - \frac{6}{14}$
 - **Similaridade entre 'deterministico' e 'determinado' é 0,4285**



SIMILARIDADE TEXTUAL

Valores agregados





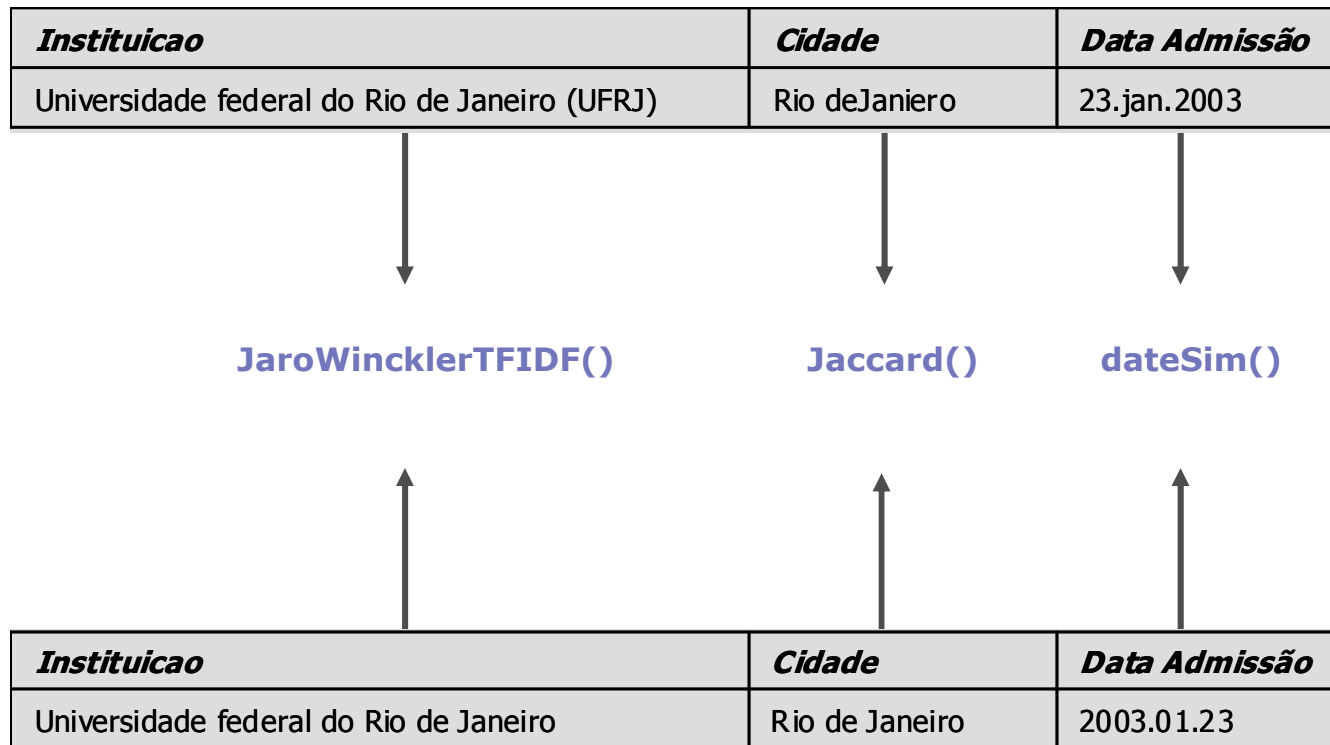
VALORES AGREGADOS

- Valores compostos por múltiplos campos
 - Tuplas, dados XML, registros
 - Funcionamento:
 - Compara cada campo individualmente, depois combina



VALORES AGREGADOS

- Idealmente
 - Cada atributo é comparado, usando uma função diferente



VALORES AGREGADOS

- Idealmente
 - Cada atributo é comparado, usando uma função diferente

<i>Instituicao</i>	<i>Cidade</i>	<i>Data Admissão</i>
Universidade federal do Rio de Janeiro (UFRJ)	Rio de Janeiro	23.jan.2003

JaroWincklerTFIDF()

Jaccard()

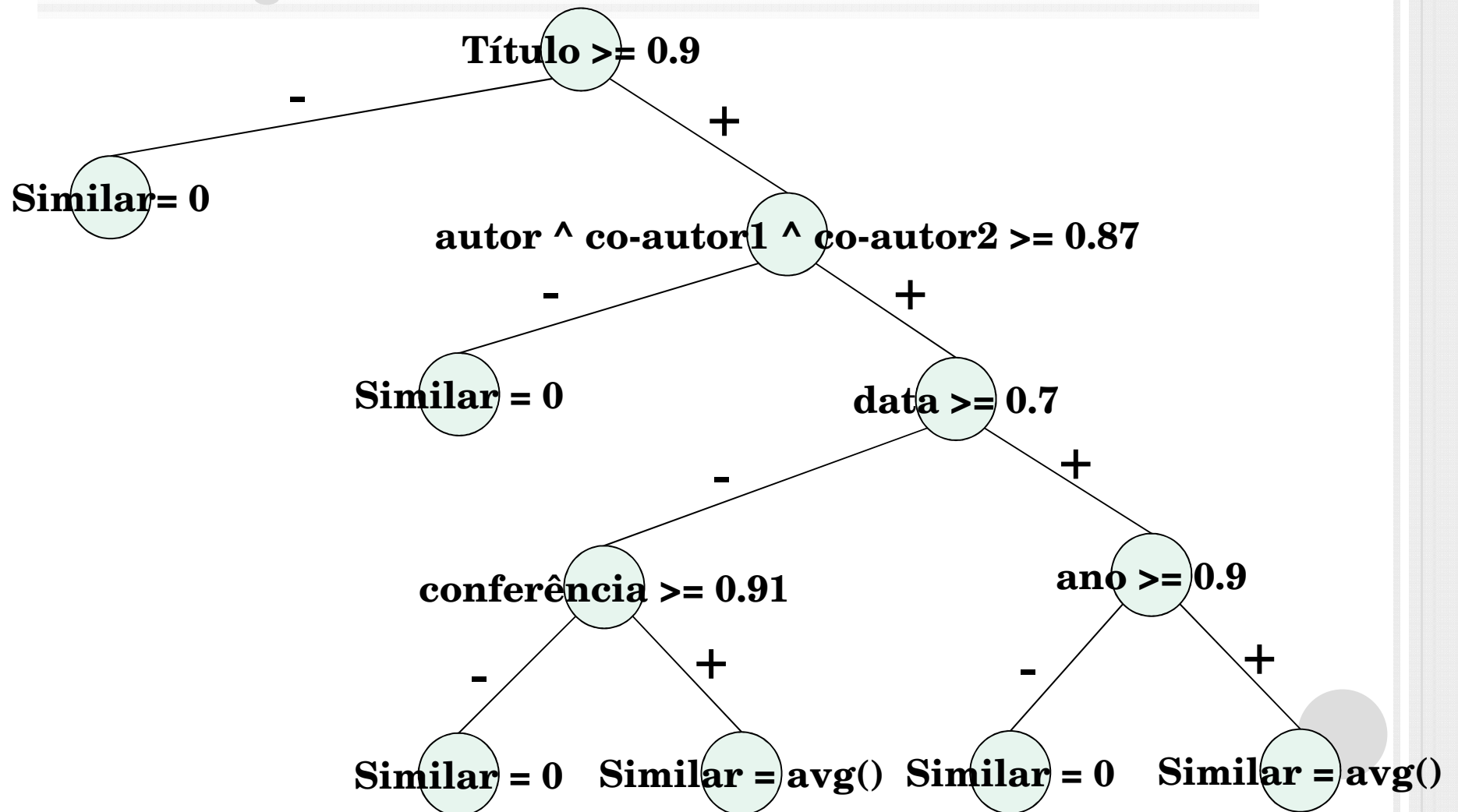
Como combinar??

<i>Instituicao</i>	<i>Cidade</i>	<i>Data Admissão</i>
Universidade federal do Rio de Janeiro	Rio de Janeiro	2003.01.23

ALGORITMOS PARA COMBINAÇÃO

- Grande parte dos trabalhos
 - Uso de algoritmos de *machine learning*
 - Árvores de decisão
 - SVM (*Support Vector Machine*)
- Outros
 - Uso de técnicas de RI
 - Combinação de rankings (*rank merge*)

ÁRVORES DE DECISÃO



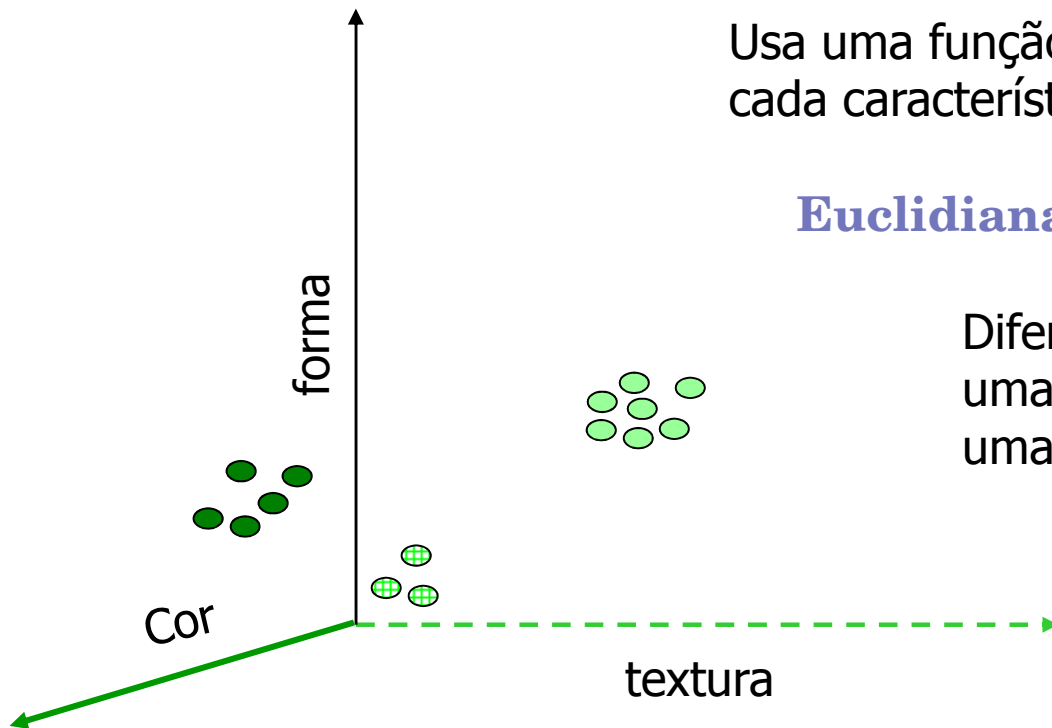


SIMILARIDADE DE IMAGENS



VETOR DE CARACTERÍSTICA

- Cria-se um espaço **n** dimensional
 - **n** é o número de características consideradas
 - Cada imagem é transformada em um vetor de características e colocada no espaço



Usa uma função para combinar os valores de cada característica

$$\text{Euclidiana } (i1, i2) = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

Diferença representa a distância entre uma determinada característica de uma imagem em relação a outra



SIMILARIDADE TEXTUAL

Textos longos



DIFF: TEXTO NÃO-ESTRUTURADO

Linha	Texto1	Texto2
1	A	W
2	B	A
3	C	B
4	D	X
5	E	Y
6	F	Z
7	G	E

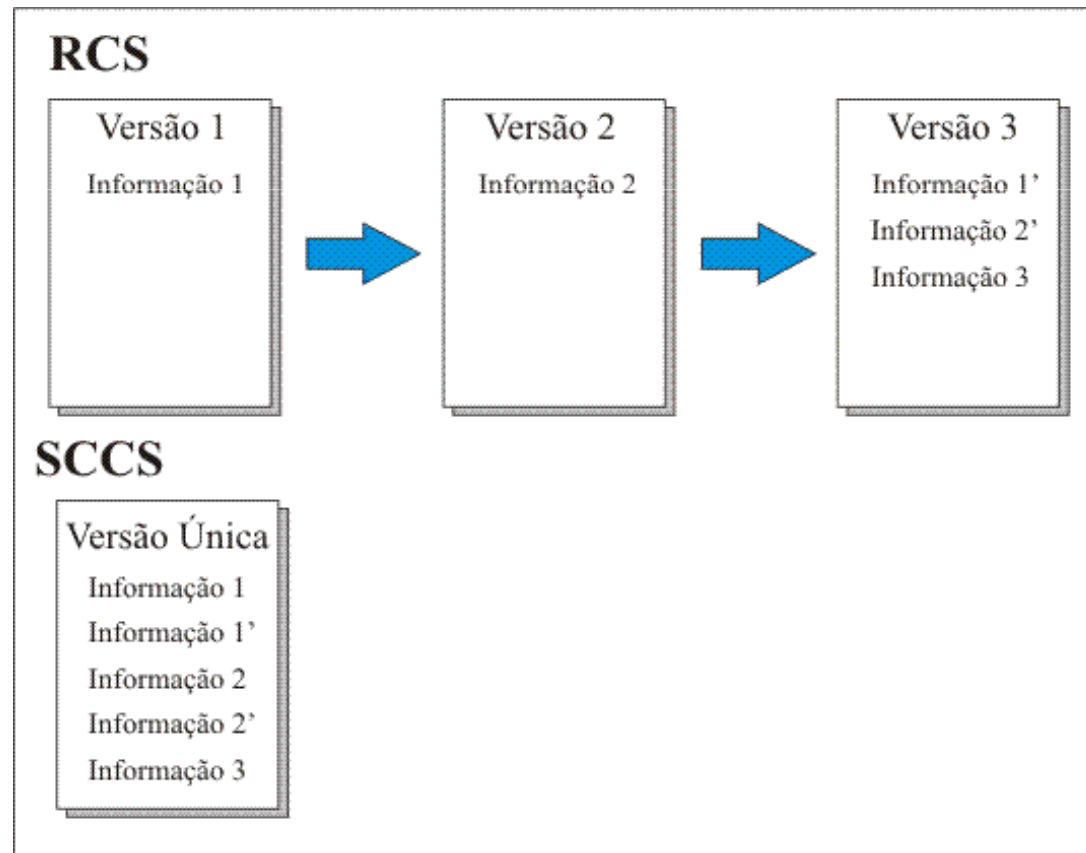
1. inserir W no início, antes da 1ª linha
2. substituir as linhas 3 e 4 (C e D) por X, Y e Z
3. deletar as linhas 6 e 7 (F e G)

DIFF: TEXTO NÃO-ESTRUTURADO

- Objetivo do algoritmo de diff
 - Reportar o número mínimo de mudanças de linhas
 - Maximizar o número de linhas deixadas inalteradas
- **LCS** (Longest Common Subsequence)
 - Base do funcionamento dos algoritmos de diff

DIFF: TEXTO NÃO-ESTRUTURADO

- SCCS (*Source Code Control System*)
- RCS (*Revision Control System*)



PROBLEMA

```
<Books>
  <book>
    <title>Harry Potter and the Sorcerer's Stone</Title>
    <Author>J.K. Rowling</Author>
    <Seller>
      <ID>Mike</ID>
      <Rating>30</Rating>
    </Seller>
    <PrimeiraOferta>$5.00</First_Bid>
    <OfertaAtual Time_Left = "36 hrs.">$8.50</Current_Bid>
    <Bidder>
      <ID>Steve</ID>
      <Rating>25</Rating>
    </Bidder>
  </Book>
  <Book>
    <Title>The Adventures of Tom Sawyer</Title>
    <Author>Mark Twain</Author>
    <Seller>
      <ID>Sean</ID>
      <Rating>100</Rating>
    </Seller>
    <First_Bid>$2.00</First_Bid>
    <Current_Bid Time_Left = "4 hrs.">$3.50</Current_Bid>
    <Bidder>
      <ID>Tim</ID>
      <Rating>5</Rating>
    </Bidder>
  </Book>
</Books>
```

PROBLEMA

```
<Books>
  <book>
    <title>Harry Potter and the Sorcerer's Stone</Title>
    <Author>J.K. Rowling</Author>
    <Seller>
      <ID>Mike</ID>
      <Rating>30</Rating>
    </Seller>
    <PrimeiraOferta>$5.00</First_Bid>
    <OfertaAtual Time_Left = "36 hrs.">$8.50</Current_Bid>
    <Bidder>
      <ID>Steve</ID>
      <Rating>25</Rating>
    </Bidder>
  </Book>
  <Book>
    <Title>The Adventures of Tom Sawyer</Title>
    <Author>Mark Twain</Author>
    <Seller>
      <ID>Sean</ID>
      <Rating>100</Rating>
    </Seller>
    <First_Bid>$2.00</First_Bid>
    <Current_Bid Time_Left = "4 hrs.">$3.50</Current_Bid>
    <Bidder>
      <ID>Tim</ID>
      <Rating>5</Rating>
    </Bidder>
  </Book>
</Books>
```


PROBLEMA

```
<Books>
  <book>
    <title>Harry Potter and the Sorcerer's Stone</Title>
    <Author>J.K. Rowling</Author>
    <Seller>
      <ID>Mike</ID>
      <Rating>30</Rating>
    </Seller>
    <PrimeiraOferta>$5.00</First_Bid>
    <OfertaAtual Time_Left = "36 hrs.">$8.50</Current_Bid>
    <Bidder>
      <ID>Steve</ID>
      <Rating>25</Rating>
    </Bidder>
  </Book>
  <Book>
    <Title>The Adventures of Tom Sawyer</Title>
    <Author>Mark Twain</Author>
    <Seller>
      <ID>Sean</ID>
      <Rating>100</Rating>
    </Seller>
    <First_Bid>$2.00</First_Bid>
    <Current_Bid Time_Left = "4 hrs.">$3.50</Current_Bid>
    <Bidder>
      <ID>Tim</ID>
      <Rating>5</Rating>
    </Bidder>
  </Book>
</Books>
```

```
<Books>
  <Book>
    <Title>The Adventures of Tom Sawyer</Title>
    <Author>Mark Twain</Author>
    <Seller>
      <ID>Sean</ID>
      <Rating>100</Rating>
    </Seller>
    <First_Bid>$2.00</First_Bid>
    <Current_Bid Time_Left = "2 hrs.">$4.50</Current_Bid>
    <Bidder>
      <ID>Tim</ID>
      <Rating>5</Rating>
    </Bidder>
  </Book>
  <Book>
    <Title>Harry Potter and the Sorcerer's Stone</Title>
    <Author>J.K. Rowling</Author>
    <Seller>
      <ID>Mike</ID>
      <Rating>30</Rating>
    </Seller>
    <First_Bid>$5.00</First_Bid>
    <Current_Bid Time_Left = "34 hrs.">$10.00</Current_Bid>
    <Bidder>
      <ID>Mark</ID>
      <Rating>125</Rating>
    </Bidder>
  </Book>
</Books>
```

PROBLEMA

```
<Books>
  <book>
    <title>Harry Potter and the Sorcerer's Stone</Title>
    <Author>J.K. Rowling</Author>
    <Seller>
      <ID>Mike</ID>
      <Rating>30</Rating>
    </Seller>
    <PrimeiraOferta>$5.00</First_Bid>
    <OfertaAtual Time_Left = "36 hrs.">$8.50</Current_Bid>
    <Bidder>
      <ID>Steve</ID>
      <Rating>25</Rating>
    </Bidder>
  </Book>
```

```
<Book>
  <Title>The Adventures of Tom Sawyer</Title>
  <Author>Mark Twain</Author>
  <Seller>
    <ID>Sean</ID>
    <Rating>100</Rating>
  </Seller>
  <First_Bid>$2.00</First_Bid>
  <Current_Bid Time_Left = "4 hrs.">$3.50</Current_Bid>
  <Bidder>
    <ID>Tim</ID>
    <Rating>5</Rating>
  </Bidder>
</Book>
```

```
<Books>
  <Book>
    <Title>The Adventures of Tom Sawyer</Title>
    <Author>Mark Twain</Author>
    <Seller>
      <ID>Sean</ID>
      <Rating>100</Rating>
    </Seller>
    <First_Bid>$2.00</First_Bid>
    <Current_Bid Time_Left = "2 hrs.">$4.50</Current_Bid>
    <Bidder>
      <ID>Tim</ID>
      <Rating>5</Rating>
    </Bidder>
  </Book>
```

```
<Book>
  <Title>Harry Potter and the Sorcerer's Stone</Title>
  <Author>J.K. Rowling</Author>
  <Seller>
    <ID>Mike</ID>
    <Rating>30</Rating>
  </Seller>
  <First_Bid>$5.00</First_Bid>
  <Current_Bid Time_Left = "34 hrs.">$10.00</Current_Bid>
  <Bidder>
    <ID>Mark</ID>
    <Rating>125</Rating>
  </Bidder>
</Book>
```

PROBLEMA

```
<Books>
<book>
  <title>Harry Potter and the Sorcerer's Stone</Title>
  <Author>J.K. Rowling</Author>
  <Seller>
    <ID>Mike</ID>
    <Rating>30</Rating>
  </Seller>
  <PrimeiraOferta>$5.00</First_Bid>
  <OfertaAtual Time_Left = "36 hrs.">$8.50</Current_Bid>
  <Bidder>
    <ID>Steve</ID>
    <Rating>25</Rating>
  </Bidder>
</Book>
<Book>
  <Title>The Adventures of Tom Sawyer</Title>
  <Author>Mark Twain</Author>
  <Seller>
    <ID>Sean</ID>
    <Rating>100</Rating>
  </Seller>
  <First_Bid>$2.00</First_Bid>
  <Current_Bid Time_Left = "4 hrs.">$3.50</Current_Bid>
  <Bidder>
    <ID>Tim</ID>
    <Rating>5</Rating>
  </Bidder>
</Book>
</Books>
```

```
<Books>
<Book>
  <Title>The Adventures of Tom Sawyer</Title>
  <Author>Mark Twain</Author>
  <Seller>
    <ID>Sean</ID>
    <Rating>100</Rating>
  </Seller>
  <First_Bid>$2.00</First_Bid>
  <Current_Bid Time_Left = "2 hrs.">$4.50</Current_Bid>
  <Bidder>
    <ID>Tim</ID>
    <Rating>5</Rating>
  </Bidder>
</Book>
<Book>
  <Title>Harry Potter and the Sorcerer's Stone</Title>
  <Author>J.K. Rowling</Author>
  <Seller>
    <ID>Mike</ID>
    <Rating>30</Rating>
  </Seller>
  <First_Bid>$5.00</First_Bid>
  <Current_Bid Time_Left = "64 hrs.">$10.00</Current_Bid>
  <Bidder>
    <ID>Mark</ID>
    <Rating>125</Rating>
  </Bidder>
</Book>
</Books>
```

PROBLEMA

```
<Books>
  <book>
    <title>Harry Potter and the Sorcerer's Stone</Title>
    <Author>J.K. Rowling</Author>
    <Seller>
      <ID>Mike</ID>
      <Rating>30</Rating>
    </Seller>
    <PrimeiraOferta>$5.00</First_Bid>
    <OfertaAtual Time_Left = "36 hrs.">$8.50</Current_Bid>
    <Bidder>
      <ID>Steve</ID>
      <Rating>25</Rating>
    </Bidder>
  </Book>
```

```
<Book>
  <Title>The Adventures of Tom Sawyer</Title>
  <Author>Mark Twain</Author>
  <Seller>
    <ID>Sean</ID>
    <Rating>100</Rating>
  </Seller>
  <First_Bid>$2.00</First_Bid>
  <Current_Bid Time_Left = "4 hrs.">$3.50</Current_Bid>
  <Bidder>
    <ID>Tim</ID>
    <Rating>5</Rating>
  </Bidder>
</Book>
</Books>
```

```
<Books>
  <Book>
    <Title>The Adventures of Tom Sawyer</Title>
    <Author>Mark Twain</Author>
    <Seller>
      <ID>Sean</ID>
      <Rating>100</Rating>
    </Seller>
    <First_Bid>$2.00</First_Bid>
    <Current_Bid Time_Left = "2 hrs.">$4.50</Current_Bid>
    <Bidder>
      <ID>Tim</ID>
      <Rating>5</Rating>
    </Bidder>
  </Book>
  <Book>
```

```
<Title>Harry Potter and the Sorcerer's Stone</Title>
<Author>J.K. Rowling</Author>
<Seller>
  <ID>Mike</ID>
  <Rating>30</Rating>
</Seller>
<First_Bid>$5.00</First_Bid>
<Current_Bid Time_Left = "34 hrs.">$10.00</Current_Bid>
<Bidder>
  <ID>Mark</ID>
  <Rating>125</Rating>
</Bidder>
</Book>
</Books>
```



APLICAÇÕES DE DATA MINING



COMO CALCULA?



Quantas transformações são necessárias para transformar o Bart na Lisa?



- Trocar tipo roupa – 1 ponto**
- Trocar cor dos olhos – 0 ponto**
- Trocar formato do cabelo – 1 ponto**
- Trocar tipo do sapato – 1 ponto**
- Trocar cor do sapato – 1 ponto**

Soma das transformações (4), dividido pelo número de transformações (trocas) (5) = 0.8

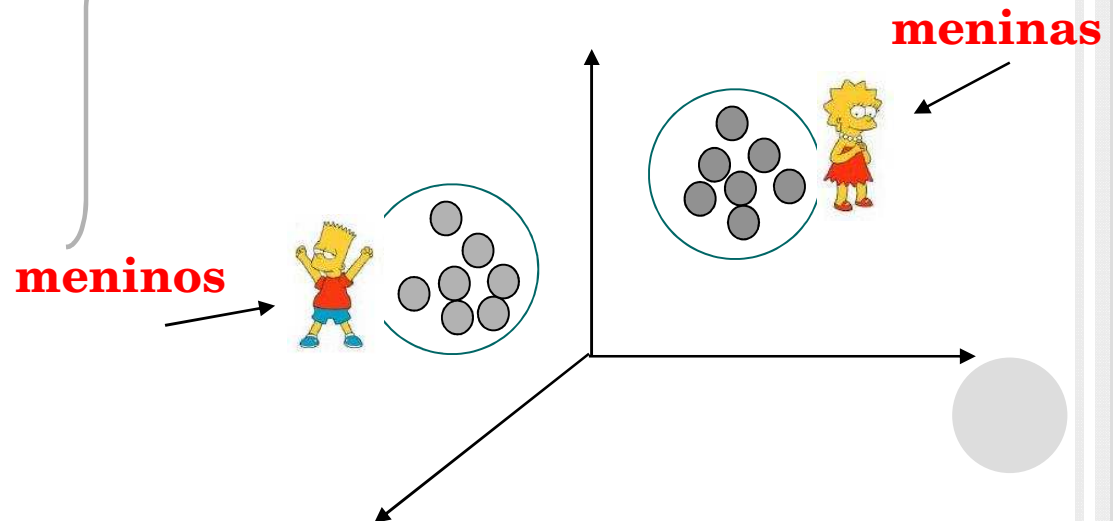
Similaridade = 1 - 0.8: 0.2

TRANSFORMAÇÕES

○ Cada transformação equivale a um atributo na tabela:

- Tipo de roupa
- Cor dos olhos
- Formato do cabelo
- Tipo de sapato
- Cor do sapato

Usando estes atributos de transformação, Bart e Lisa não pertencem ao mesmo grupo



MAS...

- Usando outros critérios de agrupamento



Quantas transformações são necessárias para transformar o Bart na Lisa?



Cliente

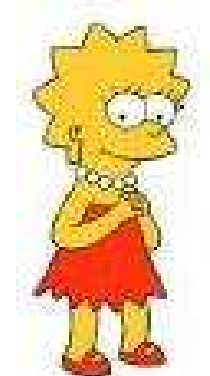
<i>nome</i>	<i>idade escolar</i>	<i>poder aquisitivo</i>	<i>nivel dep. Pais</i>	<i>ativ. Trabalho</i>	<i>ativ. Entretenimento</i>
Bart	primaria	médio	alto	nenhuma	brincar
Lisa	primaria	médio	alto	nenhuma	brincar

MAS...

- Usando outros critérios de agrupamento



Quantas transformações são necessárias para transformar o Bart na Lisa?



Trocar idade escolar – 0 ponto

Trocar poder aquisitivo – 0 ponto

Trocar nível de dependência dos pais – 0 ponto

Trocar atividade de trabalho – 0 ponto

Trocar atividade de entretenimento – 0 ponto

MAS...

- Usando outros critérios de agrupamento



Soma das transformações (0), dividido pelo número de transformações (5) = 0

Similaridade = 1 - 0: 1



Trocar idade escolar – 0 ponto

Trocar poder aquisitivo – 0 ponto

Trocar nível de dependência dos pais – 0 ponto

Trocar atividade de trabalho – 0 ponto

Trocar atividade de entretenimento – 0 ponto

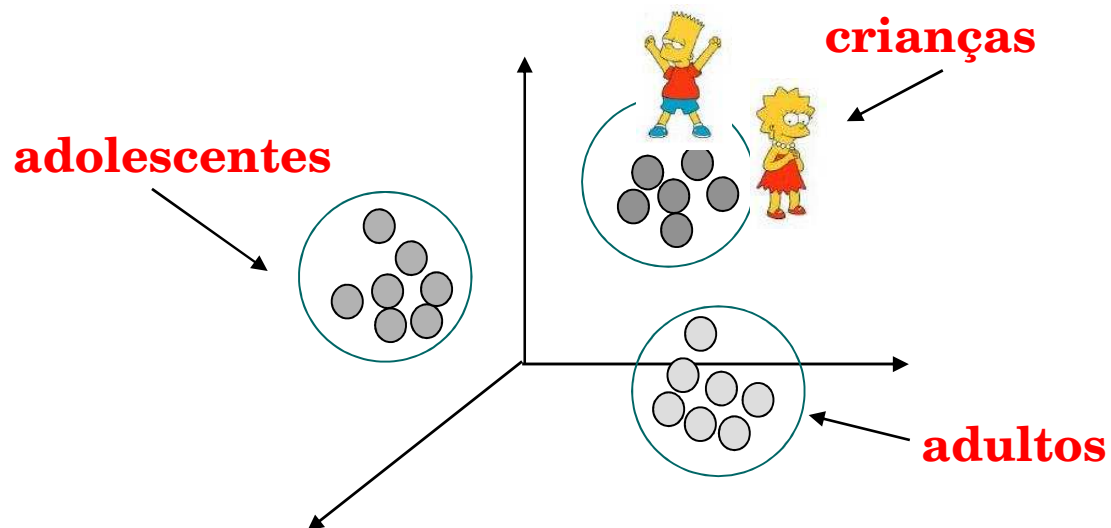
MAS...

- Usando outros critérios de agrupamento



Soma das transformações (0), dividido pelo número de transformações (5) = 0

Similaridade = 1 - 0: 1



A FUNÇÃO É BOA OU RUIM??

- Como saber se a função é boa ou ruim?
- Usa-se métricas de avaliação:
 - Revocação (ou *Recall*)
 - Precisão (ou *Precision*)
 - F-value

REVOCAÇÃO E PRECISÃO

- Avaliação é feita sobre uma base de dados conhecida
- Parâmetros
 - N = conjunto de dados, existentes no banco, que são relevantes para um dado objeto
 - R = conjunto de dados relevantes retornadas pelo sistema
 - Definir o que é relevante ou não é papel de um especialista



AVALIANDO AS FUNÇÕES DE SIMILARIDADE

PAB

AVALIAÇÃO

○ C
es

Por exemplo, considerando a
consulta:
"Porto Alegre","RS"

em nomes de cidade e

55

Cidade

nome	Estado	Relevante
Porto Alegre	RS	Sim
Bela Vista	PR	Não
P. Alegre	RS	Sim
Porto Alegre	Rio Grande do Sul	Sim
Porto Alegre	Rio G. do Sul	Sim
Belém	PA	Não
Porto Lucena	RS	Não
Belém	Pará	Não
PoA	RS	Sim
Pouso Alegre	RS	Não

AVALIÇÃO

- No re

Passo 2: PRECISÃO - para cada posição no ranking, calcular:

Número de itens relevantes naquela posição / posição

king

Consulta.

nome	Estado	Itens Relevantes	Posição	Precisão
Porto Alegre	RS	1	1	1
P. Alegre	RS	2	2	1
Porto Alegre	Rio Grande do Sul	3	3	1
Porto Alegre	Rio G. do Sul	4	4	1
Pouso Alegre	RS	4	5	0.8
PoA	RS	5	6	0.8333333333
Bela Vista	PR	5	7	0.714285714
Belém	PA	5	8	0.625
Porto Lucena	RS	5	9	0.5555555556
Belém	Pará	5	10	0.5

AVAI

o No

Passo 3: REVOCAÇÃO - para cada posição no ranking, calcular:
Número de itens relevantes naquela posição / número de relevante

Aqui são 5

Consulta: PoA

nome	Estado
Porto Alegre	RS
P. Alegre	RS
Porto Alegre	Rio Grande do Sul
Porto Alegre	Rio G. do Sul
Pouso Alegre	RS
PoA	RS
Bela Vista	PR
Belém	PA
Porto Lucena	RS
Belém	Pará

Itens Relevantes	Revocação
1	0.2
2	0.4
3	0.6
4	0.8
4	0.8
5	1
5	1
5	1
5	1
5	1

RELAÇÃO DE PRECISÃO/REVOCAÇÃO

- Para facilitar a avaliação dos resultados:
 - Gráfico que mostra a evolução da precisão em função da revocação.
 - curva de precisão e revocação

GRÁFICO DE PRECISÃO/REVOCAÇÃO

Nome	Estado	Precisão	Revocação
Porto Alegre	RS	1	0.2
P. Alegre	RS	1	0.4
Porto Alegre	Rio Grande do Sul	1	0.6
Porto Alegre	Rio G. do Sul	1	0.8
Pouso Alegre	RS	0.8	0.8
PoA	RS	0.833333333	1
Bela Vista	PR	0.714285714	1
Belém	PA	0.625	1
Porto Lucena	RS	0.555555556	1
Belém	Pará	0.5	1

59

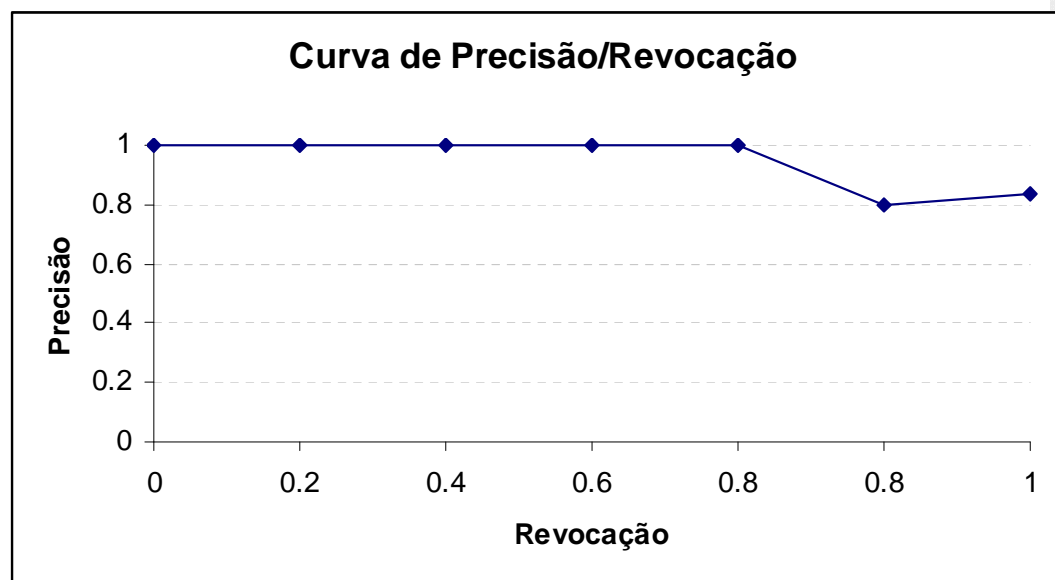
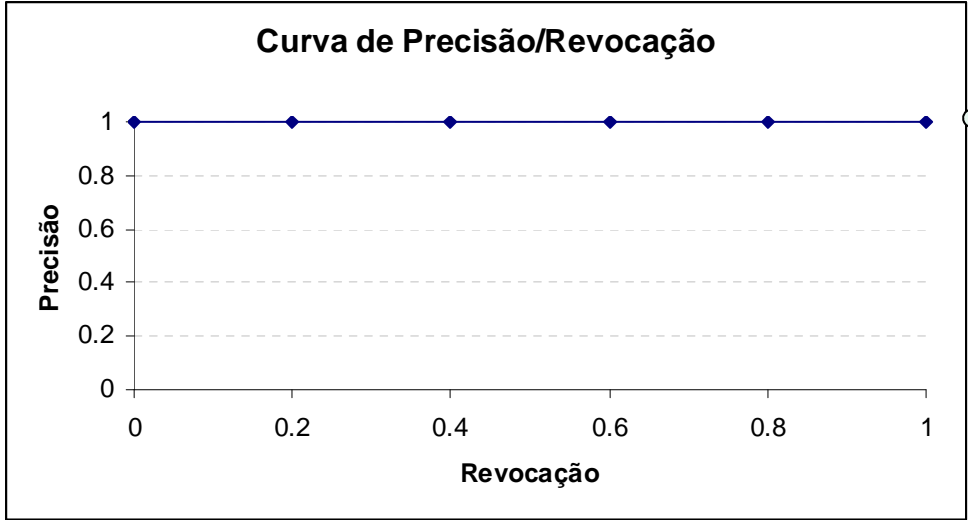


GRÁFICO DE PRECISÃO/REVOCACÃO

○ Curva ideal

Indica que todos os relevantes estão no topo do ranking



60

Nome	Estado	Precisão	Revocação
Porto Alegre	RS	1	0.2
P. Alegre	RS	1	0.4
Porto Alegre	Rio Grande do Sul	1	0.6
Porto Alegre	Rio G. do Sul	1	0.8
PoA	RS	1	1
Pouso Alegre	RS	0.833333333	1
Bela Vista	PR	0.714285714	1
Belém	PA	0.625	1
Porto Lucena	RS	0.555555556	1
Belém	Pará	0.5	1