Estudo Comparativo de Estratégias para o Pareamento de Nomes de Entidades na Língua Portuguesa

Antônio Mamede Araújo de Medeiros, **Eduardo Corrêa Gonçalves**

Escola Nacional de Ciências Estatísticas (ENCE/IBGE)

ERBD 2023

Sumário

Introdução (1/2)

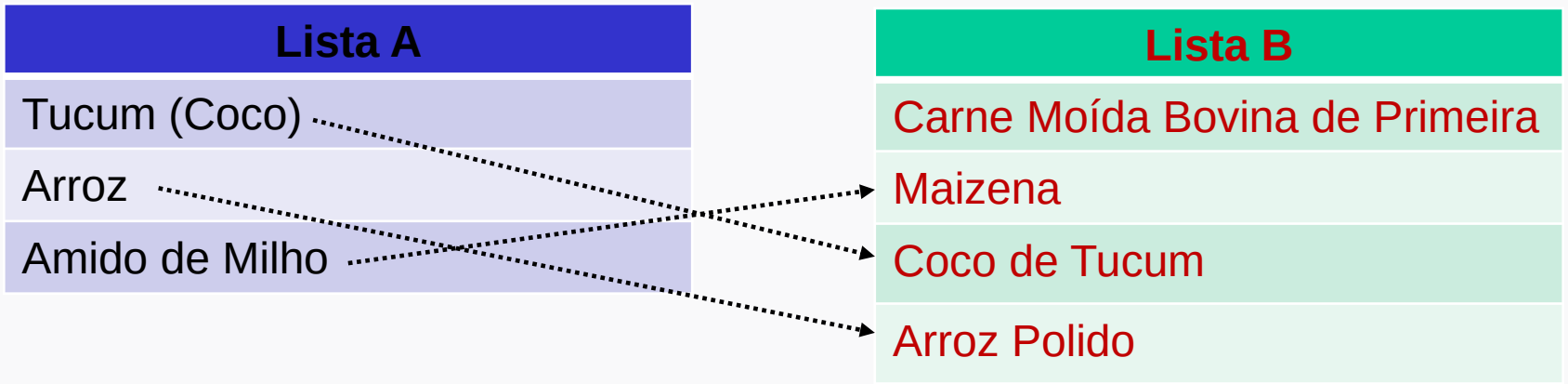
- **Q1: O que é pareamento de nomes de entidades?**
  - Tarefa que consiste em realizar a **correspondência automática de nomes (*strings*)** de uma lista A com os de uma outra lista B.
  - **Exemplo**: casar nomes de **produtos**

Lista A
Tucum (Coco)
Arroz
Amido de Milho

Lista B
Carne Moída Bovina de Primeira
Maizena
Coco de Tucum
Arroz Polido

Introdução (1/2)

- **Q1: O que é pareamento de nomes de entidades?**
  - Tarefa que consiste em realizar a **correspondência automática de nomes (*strings*)** de uma lista A com os de uma outra lista B.
  - **Exemplo**: casar nomes de **produtos**



Introdução (2/2)

- **Q2: Para que serve o pareamento de nomes de entidades?**
  - Muitas aplicações.
  - **Ex.**: **coleta automática de preços**, para identificar produtos equivalentes com o nome escrito de forma diferente.

Supermercado A

Supermercado B

Funções de Similaridade (1/2)

- **Abordagem tradicional para pareamento**
  - Usar **função de similaridade**
  - Função $S : (s_1, s_2) \rightarrow [0;1]$ que satisfaz 3 propriedades:
    1. $S(s_1, s_2) = 1$ se $s_1 = s_2$;
    2. $S(s_1, s_2) \approx 1$ quando s_1 é muito parecida com s_2, em algum sentido;
    3. $S(s_1, s_2) \approx 0$ quando s_1 é muito diferente de s_2, em algum sentido.
  - **Exemplos** de funções:
    - Levenshtein (nível alfabético)
    - Jaro-Winkler (nível alfabético)
    - Jaccard (nível léxico)
    - Cosseno de vetores TF-IDF (nível léxico)
    - Cosseno de embeddings Word2vec (nível semântico)

Funções de Similaridade (2/2)

- **Desvantagem da abordagem tradicional**
  - Isoladamente, nenhuma função consegue tratar todos os casos práticos.

$nome_1$	$nome_2$	$S_{levenshtein}$ (nível de caractere)	$S_{jaccard}$ (nível léxico)	$S_{Word2vec}$ (nível semântico)
feijão	feijwo	**0,83**	0,00	0,00
pimenta	pimentão	0,75	**0,00**	0,54
arroz com feijão	feijão com arroz	0,25	**1,00**	**1,00**
aipim	mandioca	0,25	0,00	**0,81**

Matrizes de Similaridade (1/4)

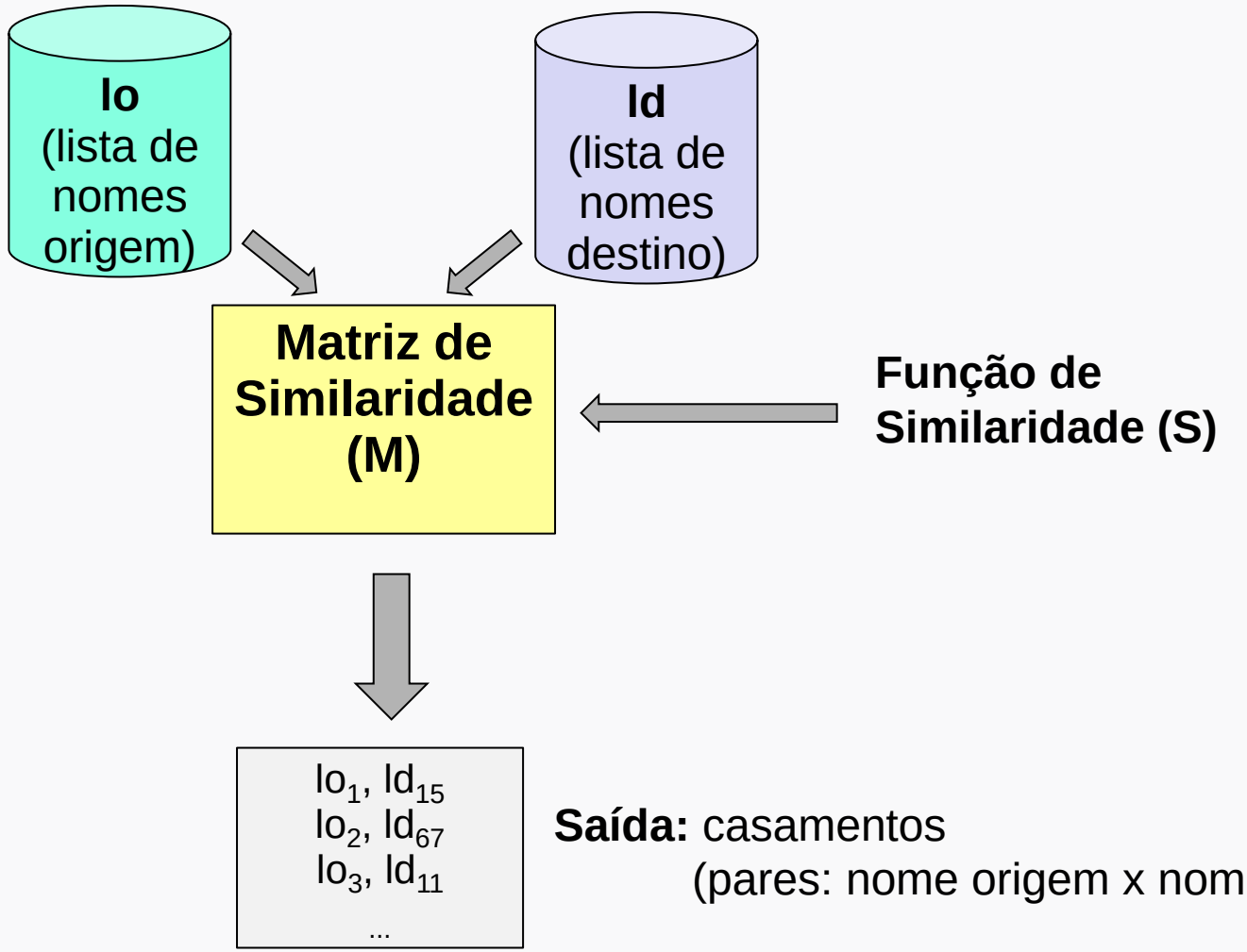
- **Solução -** Combinar funções que atuam nos diferentes níveis.
  - Alfabético
  - Léxico
  - Semântico

  - Neste trabalho, duas propostas foram combinadas e comparadas:
    - **Hartmann (2016)**
      - Originalmente criada para avaliar a similaridade de frases.
      - Combinação de TF-IDF e Word2vec.

    - **Meirelles et al. (2021)**
      - Originalmente criada para avaliar similaridade de nomes.
      - Utiliza matrizes de similaridade para permitir a combinação de *n* funções de similaridade.

Matrizes de Similaridade (2/4)

- **Matriz de Similaridade – Processo de Geração**



Matrizes de Similaridade (3/4)

- **Matriz de Similaridade - Exemplo**
  - Neste exemplo:
    - 4 nomes na lista de origem
    - 6 nomes na lista de destino
  - Os casamentos corretos são identificados pelas cores correspondentes.

		DESTINO					
		arroz	amido de milho	utensílios de plástico	atividades físicas	jogos de azar	arroz pré-cozido
ORIGEM	arroz polido	0,88	0,59	0,49	0,43	0,46	0,92
	maizena	0,56	0,40	0,41	0,51	0,47	0,47
	queijeira	0,00	0,40	0,57	0,47	0,41	0,41
	academia	0,44	0,52	0,39	0,56	0,56	0,56

Matrizes de Similaridade (4/4)

- **Matriz de Similaridade – Combinando diversas medidas**
  - Meirelles et al. (2021), propôs uma **estratégia híbrida.**
  - **Objetivo**: gerar uma matriz híbrida M_H que combina os resultados de *n* matrizes.

$$M_H(i,j) = \frac{1}{\mathbf{n}} \left(M^2_{\mathbf{1}\,(i,j)} + M^2_{\mathbf{2}(i,j)} + \cdots + M^2_{\mathbf{n}\,(i,j)} \right)$$

  - Por exemplo, pode-se combinar matrizes de:
    - Levenshtein (nível alfabético) ;
    - TF-IDF (nível léxico);
    - Word2vec (nível semântico).

  - Neste trabalho, foram comparados o desempenho de matrizes individuais e híbridas.

Experimentos (1/4)

- **Bases de Dados**
  - 3.305 pares de nomes casados de forma manual por técnicos do IBGE
    - Io = nomes da POF (Pesquisa de Orçamentos Familiares)
    - Id = nomes do SNIPC (Sistema Nacional de Índices de Preços)

Io	***Id***
ARROZ POLIDO	Arroz
ARROZ COM CASCA	Arroz
COCO BURITI	Buriti (coco)
MAIZENA	Amido de milho
QUEIJEIRA	Utensílios de Plástico
ACADEMIA	Atividades Físicas

Experimentos (2/4)

- **Metodologia p/ comparação das estratégias**

		DESTINO					
		arroz	amido de milho	utensílios de plástico	atividades físicas	jogos de azar	arroz pré-cozido
ORIGEM	arroz polido	0,88	0,59	0,49	0,43	0,46	0,92
	maizena	0,56	0,40	0,41	0,51	0,47	0,47
	queijeira	0,00	0,40	0,57	0,47	0,41	0,41
	academia	0,44	0,52	0,39	0,56	0,56	0,56

- **3 métricas**
  - **Acurácia Estrita**: proporção de vezes em que estratégia pareou corretamente o texto.
    (0 + 0 + 1 + 0) / 4 = 0,25
  - **Acurácia Ponderada**: considera resultados parcialmente corretos.
    (0 + 0 + 1 + 0,33) / 4 = 0,33
  - **Posição Média**: média do rank do par correto.
    (2 + 6 + 1 + 1) / 4 = 2,50

Experimentos (3/4)

- **Resultados – Estratégias Simples**

Matriz	Acurácia Estrita	Acurácia Ponderada	Posição Média
Levenshtein (M_L)	0,3192	0,3401	190,79
Jaro-Winkler (M_J)	0,4118	0,4127	184,43
Jaccard (M_{JC})	0,4738	0,5185	172,94
TF-IDF (M_{TF})	**0,5371**	**0,5376**	173,49
Word2vec (M_{W2V})	0,4291	0,4291	**73,83**

- Levenshtein e Jaro: desempenho pobre
  - Motivo: ruins para sinônimos, hiperônimos, ordem inversa...
- TF-IDF: melhor acurácia.
  - assim como ocorreu em Hartmann (2016)
- Word2vec: melhor posição média.

Experimentos (4/4)

- **Resultados – Estratégias Híbridas**

Matriz	Acurácia Estrita	Acurácia Ponderada	Posição Média
Levenshtein + Jaro + Jaccard + word2vec (M_{H1})	0,5322	0,5322	95,16
Levenshtein + Jaro + Jaccard + TF-IDF + Word2vec (M_{H2})	0,5625	0,5625	94,44
TF-IDF + Word2vec (M_{H3})	0,5340	0,5340	**69,55**
Jaro + TF-IDF + Word2vec (M_{H4})	**0,5673**	**0,5673**	89,26

- TF-IDF + Word2vec
  - Proposta de Hartmann (2016)
  - Piora da acurácia do TF-IDF isolado, mas melhora na posição média.

- Jaro + TF-IDF + Word2vec (melhor alfabética + melhor léxica + melhor semântica).
  - Obteve a melhor acurácia.
  - Mas a introdução do Jaro faz a posição média piorar.

Comentários Finais

- **Conclusões**
  - Resultados sugerem que combinar medidas que atuam nos 3 níveis de similaridade aumenta a eficácia do processo de pareamento.
    - Acerta 57% dos 3.305 casos propostos

- **Trabalhos Futuros**
  - Avaliar técnicas mais sofisticadas, normalmente usadas para medir a similaridade de **frases** e **documentos**.
    - Exemplo:
      - BERT
      - Redes Neurais Siamesas