# Análise do Impacto do Gerador de Conjuntos de Dados em Experimentos de Deduplicação de Dados

**Levy de Souza Silva, Mirella M. Moro**

Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

{levysouza,mirella}@dcc.ufmg.br

***Abstract.*** *Using tools to create synthetic datasets is the only solution for evaluating data duplication algorithms when real datasets are not available. However, the evaluation results may be affected by the diversity and levels of parameters available in such tools. Our goal is to verify which parameters and levels impact more on the results of deduplication experiments. Hence, we perform factorial projects on datasets created with the most used tool. Results show that two parameters explain the largest variation of results.*

***Resumo.*** *Usar ferramentas para criar dados sintéticos é a solução para avaliar algoritmos de deduplicação quando bases reais não existem. Porém, os resultados da avaliação podem ser afetados pela diversidade e quantidade de parâmetros existentes. Então, nós verificamos quais parâmetros e níveis impactam mais em experimentos de deduplicação de dados. Para tal, executamos projetos fatoriais em dados criados com a ferramenta mais utilizada. Os resultados mostram que dois parâmetros explicam a maior variação dos resultados.*

## 1. Introdução

Deduplicação de dados é a tarefa de encontrar e remover registros duplicados (representam o mesmo objeto do mundo real) em um único banco de dados; e *record linkage* e *entity resolution* em vários bancos de dados [Christen 2012]. Via de regra, abordagens desse tipo são usadas para melhorar a qualidade e a integridade dos dados. Por exemplo, em recuperação de informação, são importantes para remover documentos duplicados (páginas *web* e citações bibliográficas) retornados por motores de busca, bibliotecas digitais e sistemas de indexação automática de texto [de Carvalho et al. 2011; Hajishirzi et al. 2010]. Com a expansão de comércio eletrônico, outra aplicação em crescimento é a identificação de produtos duplicados em sistemas de lojas *online* [Christen 2012].

Várias soluções têm sido propostas visando encontrar e remover duplicatas, incluindo abordagens de programação genética [de Carvalho et al. 2012]. Independente do mecanismo proposto, um dos problemas da avaliação experimental para tais soluções é a aquisição de *datasets* reais que contenham registros originais e duplicados. Um motivo é que muitos *datasets* do mundo real contêm informações pessoais ou sigilosas, tornando improvável a distribuição pública devido a questões de privacidade e confidencialidade.

Nesse contexto, a alternativa é usar dados gerados artificialmente para validar e avaliar as abordagens. Em resumo, os registros duplicados são criados por meio da inserção de erros e modificações em um registro original. De acordo com Ioannou et al. [2013], essas modificações podem ser sintáticas, estruturais ou semânticas (erros ortográficos, permutação de caracteres, apelidos para o mesmo objeto e abreviações). Esses

*datasets* têm a vantagem que os erros são previamente conhecidos, tornando os experimentos mais controlados. Entretanto, pode ser difícil aprender de forma eficiente como usar e customizar os parâmetros disponíveis nas ferramentas de modo a construir um melhor *dataset* para as avaliações experimentais.

Nos últimos anos, pelo menos 90 estudos[1] utilizam o *Data Set Generator Program*[2] [Christen 2012, Seção 7.6] para criar conjuntos de dados com registros duplicados. Porém, nenhum desses (ou demais artigos pesquisados) avalia o impacto que as configurações dos parâmetros podem ter nos experimentos de deduplicação. Igualmente, nenhum mensura quais parâmetros e níveis têm maior impacto nos resultados experimentais, aumentando ou diminuindo a significância dos resultados.

Desse modo, os experimentos de deduplicação podem ser comprometidos devido à diversidade de parâmetros e níveis disponíveis, pois os valores de *F-Measure* podem aumentar ou diminuir de acordo com a configuração adotada. Em outras palavras, estudos que utilizam esta ferramenta criando *datasets* para validar suas abordagens podem ter resultados duvidosos. Logo, identificar quais parâmetros e níveis mais influenciam os resultados dos experimentos é uma tarefa primordial para apoiar estudos e pesquisadores que usam essa ferramenta para testar suas abordagens e metodologias. Além disso, com essas informações, os experimentos podem ser realmente mais controlados, e as abordagens podem ser avaliadas em vários cenários de configurações de parâmetros.

Nesse contexto, o objetivo desse trabalho é analisar quais parâmetros do *Data Set Generator Program* têm maior impacto nos resultados dos experimentos de deduplicação de dados. Especificamente, este estudo busca responder às seguintes questões de pesquisa: (1) Qual é a melhor configuração de parâmetros a ser adotada na criação de um *dataset* sintético? (2) Quais parâmetros têm maior influência para obter resultados insignificantes de *F-Measure*? (3) Quais parâmetros não interferem nos resultados dos experimentos? (4) Quais os níveis dos parâmetros que diminuem os resultados de *F-Measure*? e (5) Qual o percentual de impacto que as configurações dos parâmetros têm nos experimentos de deduplicação de dados? Para alcançar o objetivo, nós executamos quatro projetos fatoriais $2^k$ [Jain 1992] em um conjunto de *datasets* criados com a ferramenta.

## 2. Trabalhos Relacionados

A Tabela 1 apresenta uma amostra de estudos que usam essa ferramenta conforme coletados do Google Acadêmico e ordenados pelo número de citações. Considerando os parâmetros informados (terceira coluna da tabela), nota-se que não existe um consenso sobre quais parâmetros e configurações são importantes, pois há estudos que reportam poucos [Draisbach et al. 2012] e muitos parâmetros [de Carvalho et al. 2012]. Além disso, existem estudos que também *omitem* parâmetros [Beskales et al. 2009; Draisbach et al. 2012; Steorts et al. 2014]. Desse modo, a realização do nosso trabalho é primordial, pois busca-se descobrir quais configurações de parâmetros são realmente necessárias para criação de um *dataset* e como eles interferem nos resultados dos experimentos e avaliações de deduplicação.

O *Data Set Generator Program* é uma ferramenta que cria registros originais e duplicados de acordo com os parâmetros escolhidos pelo usuário. Essa ferramenta faz parte

[1] Encontrados no Google Acadêmico em maio de 2017

[2] `https://cs.anu.edu.au/people/Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/node70.html`

Tabela 1: Visão geral dos estudos usando a ferramenta.

| Referência | Contribuição do estudo | Configurações dos parâmetros | #citações |
|---|---|---|---|
| [Christen 2008] | Dois métodos de classificação. Um baseado na classificação de vizinhança e o outro melhora a classificação do SVM ao adicionar vetores de peso aos conjuntos de treinamento. | Original Records, Duplicate Records, Max of Change by Field, Max Duplicate by Records | 142 |
| [Draisbach et al. 2012] | Apresenta uma estratégia que adapta o tamanho da janela para aumentar a eficiência do processo de detecção de duplicatas sem reduzir a eficácia. | Distribution, Duplicate Records | 69 |
| [de Carvalho et al. 2012] | Uma abordagem de programação genética para deduplicação de registros que combina várias provas extraídas do conteúdo dos dados para encontrar uma função de deduplicação que é capaz de identificar se duas entradas são réplicas ou não. | Original Record, Duplicate Records, Max of Change by Field, Max Duplicate by Records, Distribution | 62 |
| [Beskales et al. 2009] | Um novo modelo de incerteza que codifica o espaço de possíveis reparações correspondentes a diferentes configurações de parâmetros. | Original Record, Duplicate Records, Distribution | 39 |
| [Steorts et al. 2014] | Comparação de métodos de blocagem para Record Linkage (Simple Alternatives to Blocking, Cluster-Based Blocking, LSH-Based Approaches). | Duplicate Records, Max of Change by Field | 20 |
| [Sadinle 2017] | Uma estimativa de Bayes que permite que partes incertas da correspondência não sejam resolvidas. | Type of Change, Original Records, Duplicate Records, Max of Change by Field | 4 |

do projeto *Febrl*[3] e é baseada nas ideias apresentadas por Hernández e Stolfo [1995]. A ferramenta permite criar um *dataset* contendo: (1) nomes (baseado em tabelas de frequência para nomes e sobrenomes); (2) endereços (baseado em tabelas de frequência para endereços, códigos postais, números de ruas, estados e territórios); (3) datas (como datas de nascimento); (4) números de telefone; e (5) registros identificadores. Com a ferramenta é possível definir os seguintes parâmetros: *number of original records, number of duplicate records, maximum amount of duplicate by record, maximum amount of changes by field, maximum amount of changes by instances, probability distribution (uniform, poisson or zipf ), type of change (type, ocr, phonetic, all), e number of household*s.

## 3. Experimentos e Discussões

Inicialmente, definimos quatro projetos fatoriais $2^k$ [Jain 1992][4] para medir o impacto das configurações de parâmetros na criação de um *dataset*. Para cada projeto fatorial, executamos oito experimentos com os *datasets* criados com a ferramenta e escolhemos três parâmetros e dois níveis para análise, isto é, analisamos um projeto fatorial $2^3$. Os outros parâmetros foram configurados com valores fixos.

A identificação de registros duplicados é um processo composto de várias etapas incluindo: (1) indexação, os registros recebem um valor de chave de bloco (BKv); (2) blocagem, os registros são agrupados por BKv; e (3) comparação, os registros são comparados por meio de funções de similaridade [Christen 2012]. Desse modo, utilizando a ferramenta, criamos um *dataset* para cada configuração de projeto, sobre os quais aplicamos algoritmos de deduplicação de dados. Os algoritmos usados foram: *Soundex* para o valor da chave de bloco, *Standard Blocking* para blocagem de registros e *Jaro Winkler* para a função de similaridade [Christen 2012]. Nós indexamos os registros pelo atributo *given name* e comparamos os atributos *given name* e *surname* para encontrar as duplicatas. Nós utilizamos as métricas *Precision*, *Recall* e *F-Measure* para averiguar o impacto dos parâmetros e níveis. A Tabela 2 detalha os projetos fatoriais, os parâmetros escolhidos, os níveis dos parâmetro e os valores configurados para os outros parâmetros.

A Tabela 3 agrega as porções de efeito explicadas por cada parâmetro e suas iterações em cada projeto fatorial: (a) *primeiro* a (d) *quarto*. Observe que o valor de

[3]Projeto Febrl `http://sourceforge.net/projects/febrl`

[4]O projeto fatorial serve para analisar o impacto de fatores e níveis em uma variável resposta Y.

Tabela 2: Plano de Experimentos

| Projeto Fatorial | Parâmetros Escolhidos | Níveis | Outros Parâmetros |
|---|---|---|---|
| 1 | Max Duplicate by Record | 3 e 8 | Original Records = 1.000<br>Duplicate Records = 100<br>Distribution = uniform<br>Types of Change = all<br>Households = 1 |
| | Max Change by Field | 2 e 7 | |
| | Max Change by Instance | 7 e 10 | |
| 2 | Original Records | 1.000 e 10.000 | Max Duplicate by Record = 5<br>Max Changes by Field = 5<br>Max Changes by Instances = 10<br>Type of Change = all<br>Households = 1 |
| | Duplicate Records | 30% e 70% | |
| | Probability Distribution | Uniform e Poisson | |
| 3 | Probability Distribution | Uniform e Poisson | Original Records = 1.000<br>Duplicate Records = 100<br>Max Duplicate by Record = 5<br>Max Changes by Field = 5<br>Max Changes by Instances = 10 |
| | Type of Change | Typo e Phonetic | |
| | Households | 3 e 8 | |
| 4 | Probability Distribution | Uniform e Poisson | Original Records = 1.000<br>Duplicate Records = 100<br>Max Duplicate by Record = 5<br>Max Changes by Instances = 10<br>Households = 1 |
| | Type of Change | Typo e Phonetic | |
| | Max Change by Field | 2 e 7 | |

*F-Measure* é a métrica considerada. Para computar as variações, consideramos um nível para cada parâmetro como especificado na Tabela 3 e executamos oito experimentos.

A Tabela 3(a) apresenta os resultados do primeiro projeto fatorial que considera o valor de *F-Measure* obtido variando-se os parâmetros *Max Duplicate by Record*, *Max Change by Field* e *Max Change by Instance*. Observa-se que $89.65\%$ da variação dos resultados é explicada pelo parâmetro *Max Change by Field*. Assim, a quantidade de erros que são introduzidos nos atributos dos registros duplicados contribui de forma intensa para o valor de *F-Measure* obtido nos resultados. Então, é esperado que quanto maior o número de erros introduzidos, menor a eficácia das funções de similaridade.

A Tabela 3(b) apresenta os resultados do segundo projeto fatorial que considera o valor de *F-Measure* obtido variando-se os parâmetros *Original Records*, *Duplicate Records* e *Probability Distribution*. Observa-se que $76.77\%$ da variação dos resultados é explicada pelo parâmetro *Probability Distribution* e sua iteração com os parâmetros *Original Records* e *Duplicate Records*. Assim, o parâmetro *Probability Distribution* tem uma maior influência nos valores de *F-Measure* dos resultados dos experimentos $(33, 12\%)$. Além disso, conclui-se que os resultados também são afetados pela quantidade de registros originais e registros duplicados criados, pois a iteração entre o parâmetro *Probability Distribution* e os demais é significativa (AC = $21.29\%$ e BC = $22.36\%$).

A Tabela 3(c) apresenta os resultados do terceiro projeto fatorial que considera o valor de *F-Measure* obtido variando-se os parâmetros *Probability Distribution*, *Type of Change* e *Number of Household*. Observa-se que $89.79\%$ da variação dos resultados é explicada pelo parâmetro *Type of Change*. Assim, quando o tipo de modificação é configurado para *Phonetic*, os menores valores de *F-Measure* são obtidos, como observado nas Tabelas 3(c) e 3(d). Essa função simula variações fonéticas, isto é, nomes que soam similares. A função é baseada em um arquivo que contém pares de variações fonéticas de *sub-strings*, como por exemplo, $ph \leftrightarrow f$, or $rie \leftrightarrow ry$ [Christen 2012, Seção 4.3]. Usando esta abordagem obtêm-se uma menor eficácia das funções de similaridade.

Tabela 3: Resultado dos projetos fatoriais.

**(a) Primeiro**

| Exp. | (A) Max Duplicate by Record | (B) Max Change by Field | (C) Max Change by Instance | F-Measure | Porção de efeito explicada por cada parâmetro e suas iterações |
|---|---|---|---|---|---|
| 1 | 3 | 2 | 7 | 0.6369 | A = 0.62% |
| 2 | 8 | 2 | 7 | 0.6170 | **B= 89.65%** |
| 3 | 3 | 7 | 7 | 0.6883 | C = 0.93% |
| 4 | 8 | 7 | 7 | 0.7391 | AB = 6.70% |
| 5 | 3 | 2 | 10 | 0.6174 | AC = 0.58% |
| 6 | 8 | 2 | 10 | 0.6012 | BC = 0.64% |
| 7 | 3 | 7 | 10 | 0.7037 | ABC = 0.89% |
| 8 | 8 | 7 | 10 | 0.7204 | |

**(b) Segundo**

| Exp. | (A) Original Records | (B) Duplicate Records | (C) Probability Distribution | F-Measure | Porção de efeito explicada por cada parâmetro e suas iterações |
|---|---|---|---|---|---|
| 1 | 100 | 30% | Uniform | 0.5615 | A = 0.11% |
| 2 | 1000 | 30% | Uniform | 0.5877 | B = 4.38% |
| 3 | 100 | 70% | Uniform | 0.5786 | **C = 33.12%** |
| 4 | 1000 | 70% | Uniform | 0.5914 | AB = 4.38% |
| 5 | 100 | 30% | Poisson | 0.6360 | **AC = 21.29%** |
| 6 | 1000 | 30% | Poisson | 0.5959 | **BC = 22.36%** |
| 7 | 100 | 70% | Poisson | 0.5859 | ABC = 14.37% |
| 8 | 1000 | 70% | Poisson | 0.5922 | |

**(c) Terceiro**

| Exp. | (A) Probability Distribution | (B) Type of Change | (C) Number of Households | F-Measure | Porção de efeito explicada por cada parâmetro e suas iterações |
|---|---|---|---|---|---|
| 1 | Uniform | Typo | 3 | 0.6456 | A = 0.29% |
| 2 | Poisson | Typo | 3 | 0.6623 | **B = 89.79%** |
| 3 | Uniform | Phonetic | 3 | 0.4975 | C = 0.04% |
| 4 | Poisson | Phonetic | 3 | 0.4118 | AB = 1.86% |
| 5 | Uniform | Typo | 8 | 0.6623 | AC = 3.28% |
| **6** | **Poisson** | **Typo** | **8** | **0.7485** | BC = 4.46% |
| **7** | **Uniform** | **Phonetic** | **8** | **0.3714** | ABC = 0.28% |
| 8 | Poisson | Phonetic | 8 | 0.4122 | |

**(d) Quarto**

| Exp. | (A) Probability Distribution | (B) Max Change by Field | (C) Type of Change | F-Measure | Porção de efeito explicada por cada parâmetro e suas iterações |
|---|---|---|---|---|---|
| 1 | Uniform | 2 | Typo | 0.7684 | A = 1.60% |
| 2 | Poisson | 2 | Typo | 0.6914 | B = 0.88% |
| 3 | Uniform | 7 | Typo | 0.7758 | **C = 85.18%** |
| **4** | **Poisson** | **7** | **Typo** | **0.8187** | AB = 0.58% |
| **5** | **Uniform** | **2** | **Phonetic** | **0.3741** | AC = 3.26% |
| 6 | Poisson | 2 | Phonetic | 0.5789 | BC = 1.44% |
| 7 | Uniform | 7 | Phonetic | 0.4737 | ABC = 7.07% |
| 8 | Poisson | 7 | Phonetic | 0.4627 | |

A Tabela 3(d) apresenta os resultados do quarto projeto fatorial que avalia a combinação dos parâmetros mais influentes dos projetos fatoriais um, dois e três. Observa-se que o parâmetro *Type of Change* explica $85.18\%$ da variação dos resultados.

Considerando a primeira questão de pesquisa (i.e., a melhor configuração de parâmetros), observa-se que os maiores valores de *F-Measure* são obtidos usando os níveis da Tabela 3(d). Para a segunda questão de pesquisa (i.e., parâmetros de maior influência), conclui-se que os parâmetros *Probatility Distributiton*, *Max of Change by Field* e *Type of Change* explicam uma maior variação dos resultados, sendo o último o mais influente dos três. Para a terceira questão de pesquisa (i.e., parâmetros que não interferem), observa-se que os parâmetros *Original Records*, *Duplicate Records*, *Max Change by Instance* e *Number of Households* não influenciam muito os resultados.

Por fim, para a quarta (i.e., parâmetros que diminuem os resultados de *F-Measure*) e quinta questões de pesquisa (i.e., percentual de impacto), verifica-se que a definição dos parâmetros influencia o *F-Measure*. De acordo com as Tabelas 3(c) e 3(d), o valor de *F-Measure* variou entre $0.3714$ e $0.7485$ e entre $0.3741$ e $0.8187$ respectivamente, ficando claro que os níveis dos parâmetros influenciam nos resultados dos experimentos de deduplicação. Considerando a Tabela 3(d), por exemplo, observa-se que o valor de *F-Measure* obtido teve um aumento de $135.86\%$ com o nível *Typo* ao invés do *Phonetic* para

o parâmetro *Type of Change*. Assim, estudos que não informam a configuração utilizada neste parâmetro podem ter resultados duvidosos (ex., [Steorts et al. 2014]).

## 4. Conclusões e Trabalhos Futuros

Este trabalho apresentou uma avaliação experimental do impacto dos parâmetros e níveis na criação de um *dataset* usando o *Data Set Generator Program* em experimentos de deduplicação de dados. Foram executados quatro projetos fatoriais $2^k$ em conjuntos de *datasets* criados com a ferramenta. Além disso, foram discutidas questões de pesquisas relacionadas a identificar quais parâmetros e níveis interferem nos resultados dos experimentos. Nossa principal conclusão é que os parâmetros *type of change* e *max change by field* explicam uma maior variação dos resultados. Como trabalhos futuros, planejamos realizar replicações nos experimentos, calcular o intervalo de confiança dos resultados e executar testes estatísticos para validar os experimentos.

**Agradecimentos**. Trabalho parcialmente financiado por FAPEMIG e CNPq.

## Referências

Beskales, G., Soliman, M. A., Ilyas, I. F., and Ben-David, S. (2009). Modeling and querying possible repairs in duplicate detection. *Proceedings of the VLDB Endowment*, 2(1):598–609.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *SIGKDD*, pages 151–159, Las Vegas, Nevada, USA.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, Berlin.

de Carvalho, A. P., Ferreira, A. A., Laender, A. H., and Gonçalves, M. A. (2011). Incremental unsupervised name disambiguation in cleaned digital libraries. *JIDM*, 2(3):289–304.

de Carvalho, M. G., Laender, A. H., Gonçalves, M. A., and da Silva, A. S. (2012). A genetic programming approach to record deduplication. *TKDE*, 24(3):399–412.

Draisbach, U., Naumann, F., Szott, S., and Wonneberg, O. (2012). Adaptive windows for duplicate detection. In *ICDE*, pages 1073–1083, Arlington, Virginia, USA.

Hajishirzi, H., Yih, W.-t., and Kolcz, A. (2010). Adaptive near-duplicate detection via similarity learning. In *SIGIR*, pages 419–426, Geneva, Switzerland.

Hernández, M. A. and Stolfo, S. J. (1995). The merge/purge problem for large databases. In *SIGMOD*, pages 127–138, San Jose, CA, USA.

Ioannou, E., Rassadko, N., and Velegrakis, Y. (2013). On generating benchmark data for entity matching. *Journal on Data Semantics*, 2(1):37–56.

Jain, R. (1992). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518):600–612.

Steorts, R. C., Ventura, S. L., Sadinle, M., and Fienberg, S. E. (2014). A comparison of blocking methods for record linkage. In *PSD*, pages 253–268, Ibiza, Spain.