Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Junção de dois bancos de dados com auxílio de uma variável chave utilizando o método Record Linkage

por
Carlos Eduardo Araújo Del'Isola

Brasília, 2016

Carlos Eduardo Araújo Del'Isola

Junção de dois bancos de dados com auxílio de uma variável chave utilizando o método Record Linkage

Relatório Final apresentado, como parte dos requisitos para a obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Donald Matthew Pianto

Brasília

2016

Agradecimentos

Eu dedico esse trabalho aos meus pais.

Resumo

Frequentemente, um pesquisador precisa juntar dois bancos de dados sem acesso a uma chave única, como CPF, mas tendo acesso aos nomes dos indivíduos nos dois bancos. Nomes tem o problema de não serem únicos, fato tratado na literatura usando variáveis auxiliares como sexo, nome da mãe e data de nascimento. Neste trabalho focamos no problema encontrado quando os nomes não são iguais nos dois bancos, o que pode levar à não junção de linhas que representam a mesma pessoa. Para explorar esse problema usamos dados da UnB e dados da Rais, ambos com CPF, para testar a eficácia e acurácia da junção inexata por nome. No software R usamos funções que identificam pares exatos e funções que calculam a distância entre sequências de caracteres. Nossos resultados mostram que, para o grupo de pessoas presentes nos dois bancos com nomes diferentes: 52,21 % são encontrados fazendo um pareamento exato do primeiro e último nome; 16,35 % são encontrados usando a menor distância entre o nome completo; 18,27 % são encontrados usando a menor distância entre o primeiro e último nome; e 13,17 % não são encontrados. Assim, pesquisadores que usam a metodologia de junção proposta podem ter confiança de encontrar grande parte das pessoas procuradas, mesmo quando os nomes diferem entre os dois bancos.

Palavras-chave: Record Linkage, Banco de dados, R, Função match, Função Stringdistmatrix, Junção de bancos.

Abstract

Often a researcher needs to merge two databases without access to a unique key, such as CPF, but having only access to the names of individuals in the two databases. Names have the problem of not being unique, fact treated in the literature using auxiliary variables such as sex, mother's name and date of birth. In this paper we focus on the problem encountered when the names are not the same in the two databases, which can lead to the fields from one database for a given person not being aligned with the fields for the same person from the other database. To explore this issue we use data from UNB and Rais, both containing CPF, to test the effectiveness and accuracy of inexact junction by name. In the R software use functions that identify exact pairs and functions that calculate the distance between strings. Our results show that, for the group of people present in the two banks with different names: 52.21 % are found making an exact matching of the first and last name; 16.35 % are found using the shortest distance between the full name; 18.27 % are found using the shortest distance between the first and last name; and 13.17 % are not found. Thus, researchers using the proposed joint methodology can have confidence of matching most of the people, even when the names differ between the two databases.

Keywords: Record Linkage, Databases, R, Function Match, Function Stringdistmatrix, To merge databases.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Atualmente existe uma grande necessidade de juntar bancos de dados, grande parte dessa necessidade ocorre devido a inexistência de um sistema nacional, e até mesmo em nível internacional, que agrupa os dados das pessoas em um mesmo ambiente (ou que tenha uma variável que seja capaz de identificar os indivíduos de forma única independente do banco de dados que essa pessoa estiver cadastrada), ou também quando há interesse em querer analisar dados em bancos de dados diferentes.

Infelizmente desde o surgimento da institucionalização da informatização dos dados isso não ocorreu de forma sistematizada, fazendo com que os dados ficassem dispersos, ou seja, sempre houve vários bancos de dados, que não conversam entre si, com dados das mesmas pessoas, o que acarretou no surgimento de uma gama muito grande de dados repetidos, ou quando esses dados não são repetidos, não há ligações entre os bancos, ou seja, não existe uma chave principal que proporcione uma forma de conexão entre os bancos, logo, não há uma forma de identificar quais dados de um banco se relacionam com outro banco.

O intuito desse trabalho é criar uma forma para tentar resolver esse problema tão complicado que é a junção de bancos de dados que não tem uma variável que identifique os dados presentes nos diferentes bancos de dados. Dessa forma procuramos criar um método para tentar resolver esse problema e para isso utilizaremos o método Record Linkage, pois ele traz justamente essa ideia de junção de banco de dados utilizando uma variável principal e, até mesmo outras variáveis secundárias para auxiliar a chave principal. Como há poucas variáveis que proporcionam uma identificação de dados em um banco com outros, utilizaremos o nome como chave principal, sendo essa a única chave a ser utilizada para junção de banco de dados, portanto não faremos uso de variáveis secundárias, nos limitando apenas em usar o nome. É claro que utilizar várias variáveis concomitantemente aumenta a chance em ter mais sucesso, porém em muitos bancos de dados o nome é a única opção, sendo assim, faremos utilização somente do nome. E deixamos aqui como sugestão um possível estudo futuro fazendo uso de outras variáveis auxiliares.

A seguir será apresentado alguns problemas que foram resolvidos pela utilização do método Record Linkage.

1.1 Bancos de dados e solução de problemas

São um conjunto de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas. São coleções organizadas de dados que se relacionam de forma a criar algum sentido (informação) e dar mais eficiência durante uma pesquisa ou estudo. São de vital importância para empresas e há duas décadas se tornaram a principal peça dos sistemas de informação (**O.K. Takai; I.C.Italiano; J.E. Ferreira.**, Introdução a banco de dados, 2005).

Os primeiros registros de bancos de dados foram na década de 50 e, nessa mesma década, ideias a respeito de soluções de problemas já identificados (até aquele momento) começaram a ser resolvidos (**Furtado, Gustavo**, A história dos bancos de dados, 2013).

Com o surgimento do computador e a evolução dos bancos de dados e sua visualização ampliada fez com que vários problemas começassem a surgir, logo tiveram que pensar em como solucionar esses problemas decorrentes das observações na utilização de banco de dados.

Alguns dos problemas eram:

- Identificação e eliminação de dados duplicados;
- Identificação de famílias;
- Junção de bancos de dados na área da saúde.

Identificação e eliminação de dados duplicados

Empresas muitas vezes têm a necessidade de identificar dados duplicados dentro de grandes bancos de dados. Numa população registrada, algumas chaves principais (nome, CPF) podem estar presentes em dois ou mais bancos (ou até mesmo dentro do mesmo banco), para identificar duplicatas é necessário que os campos sejam preenchidos da mesma forma, porém muitas vezes não ocorre, o que faz com que seja necessário a utilização de um outro campo dentro do banco de dados, o que às vezes não acontece também.

Dados duplicados podem surgir em diversas situações:

- Quando um grande banco de dados é atualizado e a chave principal, que é um dado adicionado a cada linha desse banco que serve como identidade, está com erro;
- Quando há junção de bancos de dados e são feitos de forma indevida.

Identificação de famílias

Não muito distante da atualidade, várias famílias migraram de várias partes do mundo refugiadas, diversas famílias foram para o mesmo país, porém para localidades diferentes. Com os registros dessas pessoas em bancos de dados foi possível identificar possíveis elos entre os nomes e, em consequência disso, identificar pessoas que pertenciam à mesma família.

Junção de bancos de dados na área da saúde

Por uma questão administrativa e estratégica, várias junções de dados de pacientes ocorreram para que houvesse a possibilidade de ações de vigilância e avaliação de serviços de saúde. Utilizando bancos de dados epidemiológicos e administrativos para identificação de desfechos, tais como: óbitos, hospitalizações, notificações de agravos e acompanhamento de doenças crônico-degenerativas - para que pudessem ser potencializados os benefícios da manutenção de sistemas de registros eletrônicos de saúde.

Além dessas três formas de vincular, analisar e relacionar bancos de dados, existem diversas outros problemas que foram resolvidos.

Todos os problemas citados tiveram a sua solução encontrada através de um método chamado Record Linkage.

1.1.1 Problema atual

Além dos problemas, como os citados anteriormente, que foram resolvidos utilizando o método Record Linkage, recentemente tem-se discutido bastante a respeito de junção de bancos de dados, pois, até então, não existe uma forma universalmente eficaz de fazê-lo.

Sabe-se que vários tipos de informações são coletados quando é feito algum cadastro. Por vezes nem todo cadastro é feito com os mesmos campos, sendo assim, são construídos diversos bancos de dados pelo país e em muitos casos com as mesmas pessoas, e geralmente precisa-se de um banco de dados apenas, porém não é uma tarefa muito simples juntá-los, visto que, em muitos casos, o campo que os une é 'nome' e, infelizmente, nem sempre os nomes são registrados de forma correta.

No Brasil, desde que as pessoas começaram a ser registradas, houve vários problemas na escrita, diversos nomes são corriqueiramente registrados de forma incorreta e, por isso, quando deparamos com alguém chamado Joao, podemos escrever João com til (ã) sem percebemos que o correto é Joao sem o til ou, por falta de atenção, escrever Joan, Juan, Joaa etc.

Com a crescente utilização da computação em todos os ramos da sociedade, surgiu um acúmulo de informações, porém nem todas estão devidamente organizadas. Além disso,

acrescenta-se o fato da problemática na diversidade, complexidade e erros nas grafias dos nomes brasileiros.

Muitas vezes há vários bancos de dados com a chave principal, geralmente o nome - como já citado, se repetindo em vários bancos e existem informações importantes em todos eles que precisam ser analisados como um todo, porém não estão agrupados em um só. Surge então a necessidade de criar uma forma de mesclá-los, a fim de ter um banco resumindo toda a informação. Entretanto, como já mencionado, por causa da complexidade dos nomes das pessoas, os mesmos podem estar escritos de forma incorreta, havendo muita dificuldade em agrupá-los. Assim, surge a ideia de trabalhar com Record Linkage, que tem a proposta de estudar formas de agrupamento desses bancos de dados minimizando o viés e tentando agrupar as chaves de forma eficaz e, portanto, transformar vários bancos de dados em apenas um.

Para isso utilizaremos as ideias propostas pelo método Record Linkage, e assim, sugeriremos alguns métodos de comparação de strings e faremos uso de algumas métricas para avaliação dos métodos aplicados. O software que nos auxiliará nesse trabalho será: R 3.2.5.

1.2 Objetivos

1.2.1 Objetivos Gerais

O uso do método Record Linkage é, geralmente, relacionado a junção, organização, otimização de banco de dados; pois a sua ideia é sempre ligar, relacionar coisas existentes com finalidade de resumir ou encontrar elos com chaves em comum entre os bancos de dados.

Atualmente uma grande preocupação é em otimizar e melhorar a junção de banco de dados, sendo esse o objetivo do trabalho.

O método Record Linkage é utilizado em várias áreas e para cada área com uma aplicação diferente, como:

- Área médica

  Intuito de organizar e analisar informações sobre a saúde com interesse de estudar e entender algumas doenças e até mesmo no auxílio à sua prevenção;

- Área de TI (Tecnologia da informação)

  Automatizar processos (fabricação) que são dependentes de informações (dados);

- Área de Banco de dados

Diferente da área de TI, apesar de também trabalhar com automatização de processos que são dependentes de informações, aqui a preocupação é com o dado em si;

- Área de estatística

  Estudo demográfico, melhoria na organização dos dados para análises etc.;

- Outras aplicações.

1.2.2 Objetivos Específicos

O objetivo desse trabalho é verificar se a variável nome é um bom parâmetro (visto que em muitos casos bancos distintos não têm chaves principais interessantes para fazer essa junção) para utilização do método Record Linkage, afim de juntar bancos de dados distintos. Para isso, utilizaremos apenas a chave principal (nome). Lembrando que esse estudo também pode ser realizado utilizando chaves secundárias para auxiliar a principal. E a forma com que faremos essa verificação é através das métricas de cálculos de distâncias.

Aqui não há interesse em juntar bancos de dados com intuito de alguma verificação em relação aos dados presentes nos bancos. Todo o estudo é feito apenas com interesse em analisar se é viável a junção de banco de dados utilizando o nome como chave principal.

Capítulo 2

Record Linkage

2.1 Record Linkage

2.1.1 Definição

Record Linkage (RL) se refere à tarefa de encontrar registros em um ou vários conjuntos de dados referentes à mesma chave ou com utilização de uma chave secundária (por exemplo: arquivos de dados, livros, sites, bancos de dados). Essa ligação é necessária quando é preciso juntar conjuntos de dados com base em entidades que podem ou não compartilhar um identificador comum (por exemplo: chave de banco de dados, URI, número de identificação nacional), no caso de não haver relação entre os conjuntos de dados, mesmo que a chave seja a mesma, pois pode haver diferenças na forma de registro ou até mesmo no local de armazenamento, estilo ou preferência.(Lifang Gu, Rohan Baxter, Deanne Vickers, and Chris Rainsford, "Record Linkage: Current Practice and Future Directions")

2.1.2 Histórico

A ideia inicial de Record Linkage se volta a Halbert L. Dunn no artigo intitulado "Record Linkage"publicado no American Journal of Public Health em 1946 (**Dunn, Halbert L.** "Record Linkage", American Journal of Public Health, 1946). Howard Borden Newcombe iniciou a teoria moderna probabilística de record linkage em 1959 no artigo in Science (**Newcombe, H. B.; J.M. Kennedy, S.J. Axford, A. P. James**. "Automatic Linkage of Vital Records". Science, 1959) que foi formalizado em 1969 por Ivan Fellegi e Alan Sunter que provou que a regra de decisão probabilística descrita por eles estava otimizada quando as variáveis fossem condicionalmente independentes. O trabalho pioneiro "A Theory For Record Linkage"ainda é lembrado por muitas aplicações atuais em Record Linkage.(**Fellegi, Ivan; Sunter, Alan.** "A Theory for Record Linkage", Journal of the American Statistical Association, 1969)

Desde o fim dos anos de 1990, várias técnicas de aprendizagem automatizadas têm sido desenvolvidas, mesmo em condições não muito favoráveis, são usadas para estimar probabilidades condicionais requeridas pela teoria de Fellegi-Sunter. Vários pesquisadores têm reportado que a condição de independência assumida por Fellegi-Sunter em seu algoritmo é frequentemente violada; entretanto, há esforços mostrados em publicações explicitando que um modelo de dependência condicional entre atributos de comparação não apresenta resultados de melhoria na qualidade do Record Linkage.

Record Linkage pode ser feito internamente sem auxílio de um computador, mas a primeira razão para se usar computadores é que reduz ou elimina a revisão manual e reproduz melhor e mais facilmente os resultados. Correspondência computacional tem sido vantajoso por permitir um supervisionamento central de processamento, melhor controle de qualidade, agilidade, consistência e melhor reprodução dos resultados. (**Winkler, William E.** "Matching and Record Linkage", U.S. Bureau of the Census, 2011)

2.1.3 Nomenclatura

Record Linkage é o termo utilizado pelos estatísticos, epidemiologistas e historiadores, entre outros, para descrever o processo de juntar registros de um banco de dados com outro que descreve a mesma entidade. Aplicações de bancos de dados refere-se a isso como "mesclar"ou "refinamento". Cientistas computacionais frequentemente se referem a isso como "correspondência de dados"ou como "problema de identificação de objeto". Outros nomes usados para descrever o mesmo conceito: "correferência, entidade, identidade, nome, resolução de registro", "Linkar", "detecção de duplicatas", "desduplicação", "correspondência de registro", "reconciliação", "identificação de objeto", "integração de dados ou informação", "resolução de entidade", "fusão".

2.1.4 Métodos

- Processamento de dados

  Record Linkage é altamente sensível à qualidade da junção dos dados, então todos os dados sob consideração, em particular os campos de chaves de identificação, devem passar por uma avaliação para verificar a qualidade dos dados antes de manter a gravação do link entre os bancos.

  Muitas chaves identificadoras para a mesma identidade podem estar presentes de forma diferente entre (e até mesmo dentro) dos dados, o que pode ser complicado e, muitas vezes, não compreendido.

- Resolução de Identidade

No processo de inteligência operacional, normalmente alimentado por um motor de resolução de identidade, através do qual as organizações podem ligar diferentes fontes de dados, tendo em vista compreender possíveis correspondências de identidade e relações não-óbvias em vários elos de dados. Ele analisa toda a informação relativa a pessoas e / ou entidades a partir de múltiplas fontes de dados, em seguida aplica-se a probabilidade de pontuação para determinar quais identidades são correspondentes e quais, se existirem, relações não-óbvias entre essas outras identidades.

- *Data Matching*

  Enquanto as soluções de resolução de entidade incluem tecnologia de correspondência de dados, as ofertas de correspondência de muitos dados não se encaixam na definição de resolução de identidade (ou entidade). Aqui estão quatro fatores que distinguem resolução de correspondência de dados, de acordo com John Talburt, diretor do Centro de Pesquisas Avançadas UALR na resolução de entidade e qualidade da Informação (**Jhon R. Talburt**, "Entity Resolution and information quality", 2011):

  1. Funciona com ambos os registros estruturados e não estruturados, que implica no processo de extração de referências quando as fontes são não estruturadas ou semiestruturadas;;
  2. Usa regras de negócios elaborados e modelos conceituais para lidar com falta, conflitos e informações corrompidas;
  3. Utiliza não correspondência quando não há uma correspondência direta;
  4. Descobre relacionamentos não-óbvios e redes de associação, ou seja, quem está associado com o que.

  Em contraste com produtos de qualidade de dados, softwares mais potentes de resolução de identidade também incluem um mecanismo de regras e processos de fluxo de trabalho, que se aplicam a inteligência de negócios para as identidades resolvidas e seus relacionamentos. Estas tecnologias avançadas tomam decisões automatizadas e o impacto no processo de negócios é em tempo real, limitando a necessidade de intervenção humana.

2.1.5 Aplicações

- Gerenciamento de dados mestres

  A maioria dos gerenciamentos de dados mestres (GDM) usam um processo de Record Linkage para identificar registros de diferentes fontes que representam a mesma entidade. Esta ligação é utilizada para criar um "registro mestre principal"que

contém os dados sem alteração, os dados reconciliados sobre a entidade. As técnicas utilizadas em GDM são as mesmas para o registro de ligação geral. GDM expande esta correspondência não só para criar um "registro mestre principal, mas para inferir relações também, ou seja, a pessoa tem o mesmo sobrenome ou semelhante e mesmo endereço ou semelhante, isto poderia implicar que eles compartilham a relação da família.

- Armazenamento de dados e inteligência de negócios

  Record Linkage desempenha um papel fundamental no armazenamento de dados e inteligência de negócios. O servidor de dados serve para combinar dados de muitos sistemas operacionais de origem diferentes em um modelo lógico de dados, que pode, então, ser subsequentemente alimentado em um sistema de inteligência de negócios para relatórios e análises. Cada fonte de sistema operacional pode ter o seu próprio método de identificação das mesmas entidades utilizadas no modelo de dados lógico, de modo que Record Linkage, entre as diferentes fontes, torna-se necessário assegurar que a informação sobre uma entidade particular em um sistema de fonte possa ser facilmente comparada com as informações sobre a mesma entidade a partir de outro sistema de origem. Padronização de dados e subsequentes Record Linkage muitas vezes é utilizado para extração, transformação e alimentação (ETA).

- Pesquisa histórica

  Record Linkage é importante para a pesquisa de histórico social na maioria dos conjuntos de dados, tais como registros de recenseamento e registros paroquiais, que foram registrados muito antes da invenção dos números de identificação nacionais. Quando fontes antigas são digitalizadas, ligação de conjuntos de dados é um pré-requisito para o estudo longitudinal. Este processo é muitas vezes ainda mais complicado pela falta de ortografia padrão de nomes, nomes de família que mudam de acordo com o local de moradia, mudança de limites administrativos, e os problemas de verificação dos dados contra outras fontes. Record Linkage foi um dos temas mais proeminentes no campo da História e da computação na década de 1980, mas desde então tem sido sujeito a menos atenção em pesquisas.

- Prática e pesquisa médica

  Record Linkage é uma ferramenta importante na criação de dados necessários para examinar a saúde das populações e do próprio sistema de saúde. Ele pode ser usado para melhorar as explorações de dados, coleta de dados, avaliação da qualidade e da divulgação de informações. As fontes de dados podem ser examinadas para eliminar registros duplicados, para identificar subnotificação e outros casos (por exemplo na contagem de população recenseada), para criar estatísticas de saúde orientada para

pessoa, e para gerar registros de doenças e sistemas de vigilância em saúde.

Alguns registros de câncer ligam várias fontes de dados (por exemplo: internações hospitalares, patologia e relatórios clínicos e registros de óbito) para gerar seus registos. Record Linkage também é usado para criar indicadores de saúde. Por exemplo, a mortalidade fetal e infantil é um indicador geral de: desenvolvimento de um país, socioeconômico, saúde pública e serviços de saúde materna e infantil. Se registros de óbitos infantis são comparadas com registros de nascimento, é possível usar variáveis de nascimento, tais como peso ao nascimento e idade gestacional, junto com os dados de mortalidade, como a causa da morte, na análise dos dados. As interações podem ajudar em estudos de acompanhamento de coortes ou outros grupos para determinar fatores como o estado vital, estatuto de residência, ou os resultados de saúde. O rastreamento é muitas vezes necessário para o acompanhamento de coortes industriais, ensaios clínicos e pesquisas longitudinais para obter a causa da morte e câncer. Um exemplo de um sistema bem sucedido e de longa data com o uso de Record Linkage na investigação médica de base populacional é o Projeto de Epidemiologia Rochester com base em Rochester, Minnesota (**St. Sauver, JL; Grossardt BR, Yawn BP, Melton LJ 3rd, Pankratz JJ, Brue SM, Rocca WA.**. "Data Resource Profile: The Rochester Epidemiology Project (REP) medical records-linkage system", Int J Epidemiol, 2012).

2.2 Método e Aplicação no trabalho

Como o objetivo do trabalho é avaliar o modelo que será criado para junção de banco de dados, utilizaremos o método Record Linkage probabilístico para fazer essas verificações.

- Record Linkage probabilístico

  As vezes chamado de fuzzy matching (também junção probabilística ou fuzzy merging no contexto de junção de bancos de dados) dado a diferente utilização do problema utilizando Record Linkage. Tem a habilidade de encontrar match ou não-match e usando essas correspondências para calcular a probabilidade de dois dados se referirem a mesma entidade. Assim, estabelecendo valores que darão uma forma de identificar o que é match e o que não é match e caso a correspondência não seja perfeita, os valores que estão entre essas probabilidades são considerados possíveis matchs. Os cálculos dessas probabilidades são feitos através de métricas de cálculos de distancias, como a métrica de Jaro-Winkler. E o intervalo de resposta é [0;1], onde 0 indica não correspondência e 1 indica correspondência perfeita, e os valores entre 0 e 1 são as possíveis correspondências (**Blakely, Tony; Salmond, Clare** "Probabilistic record linkage and a method to calculate the positive predictive value". International Journal of Epidemiology, Dezembro 2012).

2.2.1 Métodos para o cálculo das distâncias

O método Record Linkage trabalha com comparação entre *strings*, ou seja, é verificado quão próximo várias palavras estão escritas, a fim de analisar qual a distância entre elas, e, assim, poder dizer quais são as mais próximas entre si, ou seja, num grupo de palavras é possível analisar quais são as mais parecidas e mais diferentes, podendo ter resultado numa escala de 0 a 1, onde 1 indica *match* perfeito, ou seja, as palavras são iguais, ou 0 onde não há correspondência entre as palavras.

Para fazer a junção de dois bancos de dados utilizaremos algumas técnicas que são:

- Utilizar a função match do R a fim de verificar se existe alguma correspondência imediata, ou seja, se apenas utilizando essa função já é possível eliminar dados;
- Comparar o primeiro e o último nome das pessoas do primeiro grupo com as pessoas do segundo. Aqui a ideia é verificar quais têm o primeiro e o último nome iguais ou muito semelhantes;
- Utilizar a função stringdistmatrix que analisa as distâncias dos nomes, sendo possível verificar se há correspondências.

2.2.2 Distâncias

Função Match

A função match, quando aplicada, retorna um vetor indicando se existe ou não correspondência entre os vetores verificados.

Função Stringdistmatrix

A função stringdistmatrix está no pacote "stringdist", que é usada para verificar, através de distâncias comparadas entre um nome e outro de duas matrizes diferentes, quão próximos estão esses nomes. Para fazer essas verificações são utilizados alguns métodos para o cálculo de distâncias, como:

- Método Jaro: A métrica é feita calculando a quantidade de transposições entre cada um dos *strings* dados e o número de caracteres em cada *string*, será então obtido um número de 0 a 1, onde 1 indica *match* perfeito e 0 indica que não há correspondência entre os *strings* (**Jaro**, 1989, 1995).

$$dj = \begin{cases} 0 \text{ se } m = 0 \\ \frac{1}{3}(\frac{m}{|S1|} + \frac{m}{|S2|} + \frac{m-t}{m}) \text{ caso contrário} \end{cases}$$

m: Número de caracteres iguais entre os *strings*;
t: Número de transposiçoes;
S1: Quantidade de caracteres do primeiro *string*;
S2: Quantidade de caracteres do segundo *string*;
dj: Distância de Jaro.

- Método Winkler: Essa métrica utiliza o resultado obtido pela distância de Jaro acrescido do tamanho do prefixo de ambas strings e um peso e, da mesma forma, será então obtido um número de 0 a 1, onde 1 indica *match* perfeito e 0 indica que não há correspondência entre os *strings* (**Winkler**, 1990).

$$dw = \begin{cases} dj \text{ se } dj < bt \\ dj + (lp(1 - dj)) \text{ se } dj > bt \end{cases}$$

dj: Distância de Jaro;
dw: Distância Winkler;
bt: É utilizado para dizer que a distância de Jaro pode ser recalculada. Esse valor padrão é 0,7. É exatamente o que diferencia a distância Jaro da distância Winkler;
l: Tamanho comum do prefixo dos *strings*;
p: constante de ajuste, o padrao é 0,1 e não deve exceder a 0,25.

- Método Bigram: Essa métrica trabalha com separação de cada *string* em várias de tamanho dois. A forma com que é obtida a separação do *string* é juntando os caracteres subsequentes, por exemplo: CADEIRA (CA), (AD), (DE), (EI), (IR), (RA). Esse processo é feito em ambos *strings*, ou seja, nos nomes presentes em ambos bancos de dados (**Michael Collins**, 1996).

$$db = 2(\frac{bigram(x) \bigcap bigram(y)}{bigram(x) + bigram(y)})$$

db: Distância bigram;
Bigram(x): Número de combinações possíveis com dois caracteres consequentes do *string* x;
Bigram(y): Número de combinações possíveis com dois caracteres consequentes do *string* y.

- Método Edit(Levenshtein): É dada pelo número mínimo de operações necessárias para transformar um *string* no outro. Entendemos por "operações"a inserção, deleção ou substituição de um carácter. Aqui o valor obtido, quando comparado os dois objetos varia entre 0 e infinito, onde 0 indica que os objetos são os mesmos e quanto mais distante de 0 indica que os objetos não são os mesmos (**Levenshtein**, 1965).

Todos esses métodos citados têm por objetivo comparar os *strings* e comparar quão próximos estão uns dos outros.

Capítulo 3

Aplicação

3.1 Apresentação dos dados

3.1.1 Obtenção de Dados

Para realização desse trabalho serão usados os seguintes dados:

- Dados da RAIS;
- Dados de alunos matriculados na UnB.

3.1.2 RAIS

A Relação Anual de Informações Sociais (RAIS)(Ministério do Trabalho e Previdência Social) tem por objetivo o suprimento às necessidades de controle da atividade trabalhista no país, para identificação dos trabalhadores com direito ao recebimento do abono salarial. Outras funções são o provimento de dados para a elaboração de estatísticas do trabalho e a disponibilização de informações do mercado de trabalho às entidades governamentais.

Os dados coletados pela RAIS constituem expressivos insumos para atendimento das necessidades:

- Da legislação da nacionalização do trabalho;
- De controle dos registros do FGTS;
- Dos Sistemas de Arrecadação e de Concessão e Benefícios Previdenciários;
- De estudos técnicos de natureza estatística e atuarial;
- De identificação do trabalhador com direito ao abono salarial PIS/PASEP.

Para análise foram utilizados dois bancos de dados. O primeiro banco, com 1.661.663 observações, extraído da RAIS. O segundo, com 6.209 observações, extraído da UnB.

Vale ressaltar que antes de começar o processo de análise, foi feito um refinamento do banco de dados da UnB, ou seja, foram comparados todos os CPF, da UnB com os da RAIS, e criado um campo identidade (esse campo já identifica quais os nomes do banco de dados da UnB têm correspondência com os da RAIS), para que fosse possível fazer o estudo, ou seja, antes de começar todo o processo, já havia os nomes da UnB com correspondência com os nomes da RAIS pelo CPF, porém em toda a minha análise eu não fiz utilização de nenhum campo que não fosse o nome e muito menos tive acesso aos dados dos cadastrados, tanto RAIS quanto UnB. Logo, o banco de dados da UnB era maior, e com esse filtro por CPF, houve uma diminuição para 6.209 nomes. Sendo assim, já era possível saber que haveria 6.209 ımatchs.

A intenção desse estudo é para avaliar se justifica fazer junção de banco de dados utilizando somente o nome ou não.

Outra observação a ser feita é com relação a função stringdistmatrix, pois diferente da função match, ela não traz uma resposta única, não no modo com que ela foi utilizada nesse trabalho. Quando utilizada a função match a resposta é 'há correspondência' ou 'não há correspondência', já a função stringdistmatrix necessita da escolha de uma das métricas abordadas no capítulo anterior, que são Jaro-Winkler, Bigram, Edit. Assim, a forma com que escolhemos para obter uma resposta satisfatória, ou seja, *match*, foi aplicar a função três vezes (uma vez para cada métrica) e escolher o valor máximo para cada vez que foi utilizada a função stringdistmatrix para cada nome da UnB com relação aos nomes presentes no banco de dados da RAIS e feito uma comparação entre as três respostas obtidas, caso as respostas fossem as mesmas, a correspondência era confirmada. Caso as respostas fossem diferentes, era considerado um nome sem correspondência.

3.2 Ordem dos métodos

Os métodos utilizados foram na ordem:

- Função match com os nomes completos;
- Função match com o primeiro e o último nome;
- Função stringdistmatrix com o primeiro e último nome;
- Função stringdistmatrix com o nome completo.

3.3 Análise dos resultados

3.3.1 Função match com os nomes completos

- Nomes com correspondência: 5.169 (83,25 %);
- Nomes sem correspondência: 1.040 (16,75 %);
- Nomes com mais de uma correspondência: 319 homônimos (6,17 %).

Tabela 3.1: Exemplo de nomes com match

Nome RAIS	Nome Unb
ALAN MAX SILVA NUNES	ALAN MAX SILVA NUNES
PALOMA PIORNO BALTORE	PALOMA PIORNO BALTORE
ERICK SOARES LINS	ERICK SOARES LINS

Antes de partir para o próximo passo, houve um trabalho de identificação de possíveis problemas que não foram resolvidos utilizando a função match para o nome completo, que são:

- Abreviação de nomes; (Nomes com abreviações);
- Erro na digitação (Nomes faltando letra (s), com letra (s) a mais, letra (s) errada (s));
- Nomes incompletos (Falta um ou mais nomes);
- Nomes diferentes (Nomes com um grau relevante de diferença entre os bancos de dados).

Tabela 3.2: Problemas com a função match para o nome completo

Nome RAIS	Nome Unb
Exemplo problema pela abreviação de nomes	
FELIPE L P PINHEIRO	FELIPE LUIS PEREIRA PINHEIRO
FILLIPPI AUGUSTO O SANTOS	FILLIPPI AUGUSTO OLIVEIRA DOS SANTOS
HEIBBE CRISTHIAN B OLIVEIRA	HEIBBE CRISTHIAN BENEDITO DE OLIVEIRA
Exemplo problema por erro de digitação	
THIAGO GOULART MORA	THIAGO GOULART MOURA
EDNA BARBOSA DA SILVA	EDNA SILVA BARBOSA
THOMAS MAILLEUX SANTANA	THOMAS MAILLEUX SANT' ANA
Exemplo problema por nomes incompletos	
RANIELLE NOLETO PAZ ARAUJO	RANIELLE NOLETO PAZ
CAMILA MARINHO SILVA SOUSA	CAMILA MARINHO SILVA
JEAN DAISY CORTEZ DA SILVA	JEAN DAISY CORTEZ DA SILVA NOBRE
Exemplo problema por nomes diferentes	
ANDRE LUIS DA SILVA PIMENTA	ANDRE LUIS MOTTA PIMENTA
THAIS WEIL NADER MOTTA	THAIS WEIL DA COSTA
KATHYANNE SANTOS C RODRIGUES	KATHYANNE DOS SANTOS COSTA

Tabela 3.3: Frequência de erros no match

Tipo de problema	Frequência
Abreviação de nomes	362
Erro na digitação	211
Nomes incompletos	362
Nomes diferentes	105
Total	1040

3.3.2 Função match com o primeiro e o último nome

- Nomes com correspondência: 543 (51,21 %);
- Nomes sem correspondência: 497 (48,79 %);

Tabela 3.4: Exemplo de problema com os matchs

Nome RAIS	Nome Unb
PEDRO HENRIQUE M ALBUQUERQUE	PEDRO HENRIQUE MELO ALBUQUERQUE
SOLANGE CRISTINA R FERNANDES	SOLANGE CRISTINA REGO FERNANDES
MONICA G COELHO DE CARVALHO	MONICA GABRIELLA COELHO DE CARVALHO

66,67 % eram de casos com nomes abreviados e os outros 33,33 % eram por erro de digitação no nome do meio.

3.3.3 Função stringdistmatrix com o primeiro e último nome

- Nomes com correspondência: 190 (38,23 %);
- Nomes sem correspondência: 307 (61,77 %);

Aqui identificamos os problemas por não termos conseguido usar a função "match"que foram resolvidos utilizando a função stringdistmatrix (para o primeiro e último nome), que são:

- Letra a mais;
- Letra diferente;
- Letra a menos;
- Ordem trocada de letras.

Tabela 3.5: Problemas com a função stringdistmatrix com o primeiro e último nome

Nome RAIS	Nome Unb
Exemplo problema por letra a mais	
FREDERICO DALMEIDA	FREDERICO ALMEIDA
Exemplo problema por letra diferente	
PAULA SHIMEBUKO	PAULA SHIMABUKO
GABRIAL LIMA	GABRIEL LIMA
FILIPE CARNEIRO	FELIPE CARNEIRO
Exemplo problema por letra a menos	
GUILHERME LUNA	GUILHERME LUN
EMANUELE ALBUQUERQUE	EMANUELE ALBUQUERQU
MARIA FILGUEIRAS	MARIA FILGUEIRA
Exemplo problema por ordem trocada de letra	
HEBERTH ALVES	HERBETH ALVES
MARCOS OLIVIERA	MARCOS OLIVEIRA

Tabela 3.6: Frequência de problemas com a função stringdistmatrix

Tipo de problema	Frequência
Letra a mais	1
Letra diferente	9
Letra a menos	178
Ordem trocada de letras	2
Total	190

3.3.4 Função stringdistmatrix com o nome completo

- Nomes com correspondência: 170 (38,23 %);
- Nomes sem correspondência: 137 (61,77 %);

Aqui identificamos os problemas por não termos conseguido usar a função stringdistmatrix (para o primeiro e último nome) que foram resolvidos utilizando a função stringdistmatrix (para o nome completo), que são:

- Ausência do último nome;
- Diferença no último nome;
- Diferença no primeiro nome;
- Abreviação no último nome.

Tabela 3.7: Problemas com a função stringdistmatrix para o nome completo

Nome RAIS	Nome Unb
Exemplo problema por ausência do último nome	
RANIELLE NOLETO PAZ	RANIELLE NOLETO PAZ ARAUJO
PRISCILLA GUIMARAES DE PAULA	PRISCILLA GUIMARAES DE PAULA GURGEL
CAMILA MARINHO SILVA	CAMILA MARINHO SILVA SOUSA
Exemplo problema por diferença no último nome	
THAIS WEIL NADER MOTTA	THAIS WEIL DA COSTA
MANAIRA DE PAULA ARAUJO SALES	MANAIRA DE PAULA GUEDES ARAUJO
LUCIANA BURLE GRIPP AMARAL	LUCIANA BURLE GRIPP COTTA
Exemplo problema por diferença no primeiro nome	
JAQUES GOMES DE JESUS	JAQUELINE GOMES DE JESUS
Exemplo problema por abreviação no último nome	
EMIDIO VASCONCELOS MONTEIRO JR	EMIDIO VASCONCELOS MONTEIRO JUNIOR
MOZANIEL MEDEIROS SANTOS JR	MOZANIEL MEDEIROS DOS SANTOS JUNIOR
JEFFERSON DOS SANTOS MOTTA JR	JEFFERSON DOS SANTOS MOTTA JUNIOR

Tabela 3.8: Frequência de problemas com a função stringdistmatrix

Tipo de problema	Frequência
Ausência do último nome	138
Diferença no último nome	23
Diferença no primeiro nome	1
Abreviação no último nome	8
Total	170

Depois de feito todos os procedimentos acima restaram apenas 137 nomes sem *matchs*, o que mostra que apenas 2,21 % da amostra ficaram sem correspondência.

Tabela 3.9: Exemplos de casos sem match

Nome RAIS	Nome Unb
ADRIANA GONCALVES HUNOF	ADRIANA BARBOSA GONCALVES
ANGELA GLORIA ALEXANDRE	ANGELA DOS ANJOS GLORIA
VANESSA TEREZINHA ALVES TENTES	VANESSA T A TENTES DE OUROFI
FERNANDO ANTONIO RODRIGUES DIAS	ADRIANE TOMAZELLI DIAS
IDALI FLORENCIO DA SILVA VIEIRA	GUILHERME ALEXANDRE VIEIRA

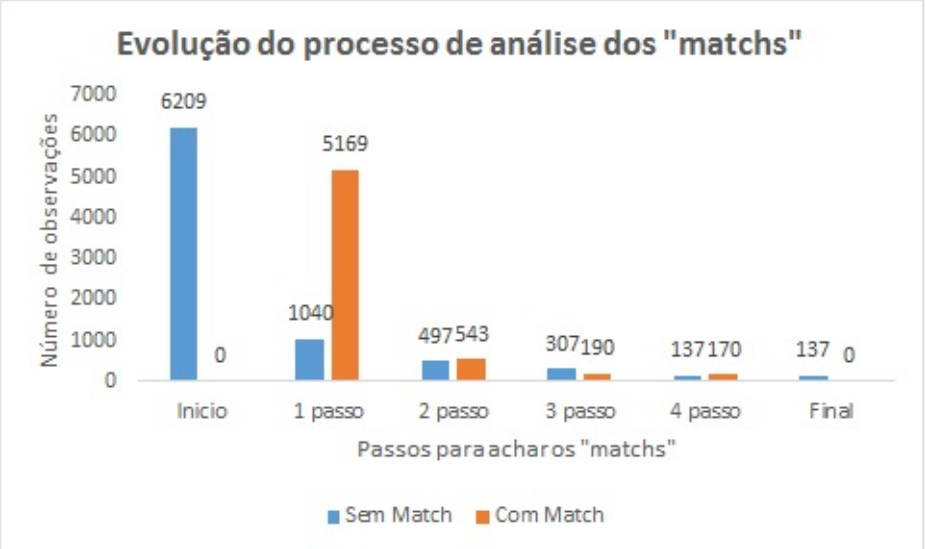

Figura 3.1: Resumo 1 análise dos matchs

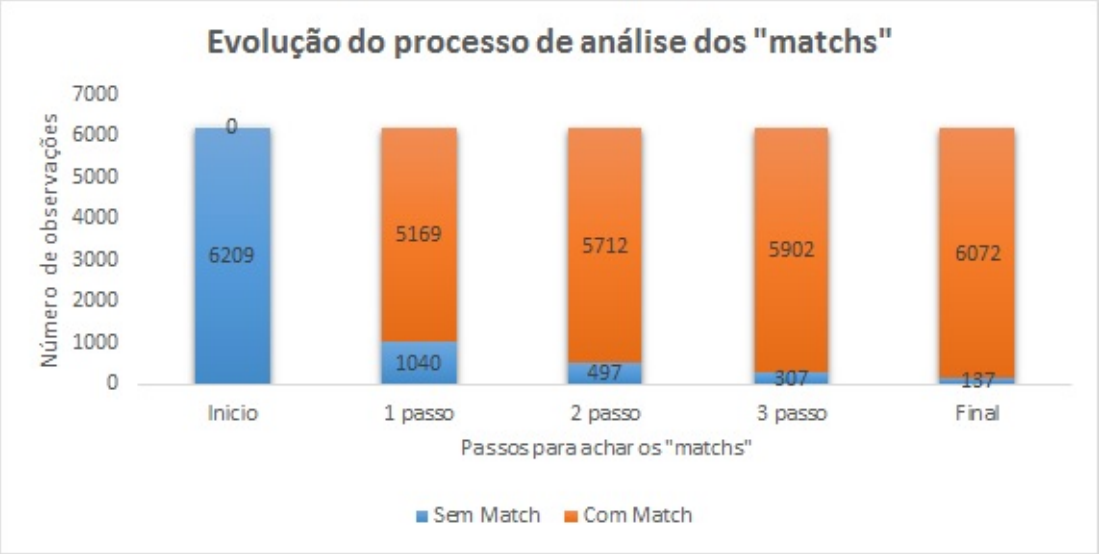

Figura 3.2: Resumo 2 análise dos matchs

3.4 Considerações Finais

Aplicando a função match nos nomes completos da UnB e da RAIS, foi possível encontrar 5.169 correspondências, sobrando 1.040 nomes no banco de dados da UnB sem correspondência, logo é possível resolver grande parte dos problemas de junção de banco de dados, utilizando apenas o nome. Entretanto houve 319 correspondências mútuas, ou seja, dos 5.169 *matchs* 319 tiveram mais de uma resposta.

Realizando o segundo procedimento (match com os nomes fragmentados em primeiro e último) para encontrar as correspondências, obtivemos 543 respostas positivas e 497 nomes sem correspondência.

No terceiro passo, com a aplicação da função stringdistmatrix reduzimos de 497 nomes para 307, ou seja, conseguimos encontrar outras 190 correspondências. É preciso salientar que a função stringdistmatrix é mais sensível ao encontrar *matchs* uma vez que ela pode obter várias respostas com o mesmo resultado, entretanto, de acordo com o parâmetro criado para chegar no resultado da comparação do nome presente no banco de dados da UnB com o nome presente no banco de dados da RAIS, não tivemos nenhuma correspondência mútua, ou seja, todos os resultados foram únicos.

Dessa vez, com a aplicação da função stringdistmatrix para o nome completo obtivemos uma resposta de 170 ımatchs, o que reduz de 307 casos para 137.

E por último, com os 137 nomes, não foi possível reduzir mais a quantidade de nomes do banco de dados da UnB sem correspondências, uma vez que esses nomes são muito diferentes dos nomes presentes no banco de dados da RAIS.

É interessante ressaltar que foram utilizados dois bancos de dados reais e um deles,

RAIS, com um número expressivo de observações, o que ocasionou alguns problemas, principalmente quando tentamos aplicar a função stringdistmatrix, pois ela traz como resposta uma matriz e o computador utilizado não suportava o tamanho da resposta, dessa forma foi preciso trabalhar com o banco de dados de diversas formas para que o computador pudesse rodar a programa afim de obter as respostas do método.

Outro problema encontrado foi no tempo para executar cada passo, exatamente pelo tamanho do banco de dados da RAIS, pois além de demorar muito, por vezes trazia a resposta que havia falta de memória. As respostas que precisava só foram possíveis obter fazendo a análise em blocos de nomes quando utilizada a função stringdistmatrix. Logo, o hardware foi um fator limitante, e nesse ponto do trabalho a limitação foi em usar a função stringdistmatrix comparando 497 nomes do banco de dados da UnB com 1.661.663, ou seja, 8GB de RAM não consegue armazenar uma matriz de resposta para 497 nomes, o que dizer de um estudo com um banco de dados maior?!

3.4.1 Conclusão

Através desse estudo percebe-se que há viabilidade em fazer junção de bancos de dados pela utilização da variável nome, pois houve 6.072 correspondências, ou seja, 97,79% dos nomes presentes no banco de dados da UnB, que antecedente a esse estudo já haviam sido relacionados com o banco de dados da RAIS, tiveram correspondência com o banco de dados da RAIS, seja essa correspondência feita de forma direta através da função "match"ou de forma indireta através da função "stringdistmatrix"ou, ainda mesmo, através da manipulação do banco de dados.

No fim, acredito que usar o nome para mesclar dois bancos de dados é uma opção razoável, visto que foi possível achar correspondência em quase todos os nomes, mas é notória a necessidade de uma segunda variável (variável auxiliar) para se ter um melhor ajustamento, pois com essa (s) variável (eis) auxiliar (es) poderíamos ter eliminado os problemas iniciais que tivemos com a função match, uma vez que, não teríamos problemas com múltiplas correspondências, porque poderíamos ter uma forma de distingui-las. Logo, este estudo torna-se mais eficaz se for utilizado, além do nome, outras variáveis em comum em ambos bancos de dados.

Referências Bibliográficas

[1] JIYOUNG SHIN. Comparative study on blocking methods in Record Linkage. **Oklahoma State University**, Estados Unidos, 2009.

[2] Nishand K.; Ramasami.S; T. Rajendran. An Efficient way of Record Linkage System and Deduplication using Indexing techniques, Classification and FEBRL Framework. **International Journal of Emerging Science and Engineering**, Índia, 2013.

[3] Rohan Baxter; Peter Christen; Tim Churches. A Comparison of Fast Blocking Methods for Record Linkage. **Australia National University**, Australia, 2003.

[4] Thomas H. Herzog; Fritz Scheuren; William E. Winkler Record linkage. **John Wiley e Sons, Inc.**, Estados Unidos, 2010.

[5] WINKLER,William E. Record Linkage Software and Methods for Merging Administrative Lists. **Bureau of the census statistical research division**, Washington D.C., 2001.

[6] WINKLER,William E. Matching and Record Linkage. **Bureau of the census statistical research division**, Washington D.C.