Identificação de relacionamento entre bases de dados através da aplicação de mineração de dados no apoio à integração da cadeia de suprimentos

Renata Moreira Limiro, Núbia Rosa da Silva, Douglas Farias Cordeiro

Resumo A integração de dados em cadeia de suprimentos é uma área que demanda especial atenção no âmbito da indústrias, apresentando demandas e desafios tais como a identificação de relacionamento entre bases de dados de diferentes contextos, as quais não possuem atributos que possibilitem uma integração direta. No que se refere à entidades de produtos de indústrias e canais de distribuição, uma das formas de se estabelecer os relacionamentos é a partir de análises sobre as descrições de produtos, o que é comumente realizado de forma manual por analistas de dados. Neste artigo é apresentada uma solução inteligente para integração de dados da cadeia de suprimentos através da aplicação de um método de mineração de dados baseado no uso de distância de Levenshtein. A avaliação do método é feita através de um conjunto de dados reais de indústria e canais de distribuição do setor do agronegócio. Os resultados demonstram a efetividade do método, com uma taxa de identificação de relacionamentos superior a 80% dos registros.

1 Introdução

A integração de dados tem sido apresentada como solução para diversos problemas relacionados a eficiência e eficácia de organizações e seus sistemas que apoiam a gestão e a tomada de decisão. Onde antes eram utilizados diversos sistemas dedicados para controlar, de forma individual e separada, diferentes departamentos de uma mesma empresa, entidade ou indústria, modelos de sistema mais integrados, como ERPs (do inglês, *Enterprise Resource Planning*), ganharam espaço, proporcionando vantagens estratégicas através da integração de dados e, consequentemente, de uma visão holística e detalhada (SÊMOLA, 2014). A partir do sucesso e da expansão de modelos de sistemas integrados, e da crescente disponibilidade de dados originados de fontes diversificadas, impulsionada pelo fenômeno do Big Data, é possível perceber a necessidade da correlação assertiva de dados advindos de diferentes contextos ou de cenários que envolvem distintos agentes, porém, que possuem relações ou impacto entre si. A cadeia de

Renata Moreira Limiro
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: renatamlimiro@gmail.com

Núbia Rosa da Silva
Instituto de Biotecnologia - Universidade Federal de Catalão, Catalão, Goiás, Brasil.
e-mail: nubia@ufg.br

Douglas Farias Cordeiro
Faculdade de Informação e Comunicação - Universidade Federal de Goiás, Goiânia, Goiás, Brasil.
e-mail: cordeiro@ufg.br

Anais do XV Encontro Anual de Ciência da Computação (EnAComp 2020). ISSN: 2178-6992.
Catalão, Goiás, Brasil. 25 a 27 de Novembro de 2020.

suprimentos (*Supply Chain*), pode ser considerada um exemplo de cenário que depende amplamente da integração eficiente de seus diferentes agentes para o alcance de melhores resultados e é nesse sentido que a Gestão da Cadeia de Suprimento (*Supply Chain Management*) se preocupa em buscar os mais diversos e avançados meios e métodos, inclusive aqueles relacionados à tecnologias inteligentes, que resultem em vantagem competitivas para os seus agentes (BOWERSOX et al., 2014).

Segundo Simchi-Levi, Kaminsky e Simchi-Levi (2010), a cadeia de suprimentos pode ser considerada como o processo no qual matérias primas são manufaturadas em fábricas e transformadas em produtos, transportadas para depósitos e armazéns, e por fim entregues para varejistas e clientes finais. Desta maneira, é possível entender que existem diferentes agentes envolvidos para que a cadeia de suprimentos funcione, desde fornecedores de matérias-primas até aqueles que possuem atividades que relacionam à entrega do produto ao cliente final. Neste sentido, ainda segundo (SIMCHI-LEVI; KAMINSKY; SIMCHI-LEVI, 2010), a Gestão da Cadeia de Suprimentos aplica seus esforços para garantir e aprimorar a integração entre os seus agentes, visando minimização dos custos envolvidos nesta cadeia e cumprindo com todas as necessidades em termos de entregas a serem realizadas.

A partir do desenvolvimento tecnológico e da evolução nos processos de vendas, manufatura, logística, armazenagem e distribuição, a Tecnologia da Informação e Comunicação tem ganhado cada vez mais espaço como ferramenta de auxílio à gestão da cadeia de suprimentos (GOMES; RIBEIRO, 2013). Neste contexto, muitos dados são gerados em resultados dos processos e de suas atividades dentro da cadeia de suprimentos, o que gera uma grande necessidade de gestão e integração para que informações úteis e de qualidade possam ser aplicadas à tomada de decisão estratégica, se levando em conta os diferentes níveis da cadeia.

Ao se tratar de diferentes agentes relacionados na cadeia de suprimentos, ou seja, do contexto do fornecimento das matérias-primas até os canais que distribuem os produtos manufaturados, se deve considerar a presença dos seguintes agentes: fornecedor da matérias-primas, transportador, indústria, centros de armazenagem e canais de distribuição (BALLOU, 2007). Dessa maneira, podem existir diferentes sistemas e diferentes culturas de gestão de dados e informações. Essas diferenças podem causar impactos na integração da cadeia, o que consequentemente gera demandas para utilização de tecnologias inteligentes que permitam integração das informações com visibilidade para os distintos agentes da cadeia de suprimentos.

Considerando os agentes indústria e seus canais de distribuição, é interessante observar que ambos se relacionam através de transações de um produto manufaturado. No entanto, cada um possui uma visão diferente, a indústria acompanha a manufatura e o *sell in* (indicador que mensura os produtos vendidos da indústria para o distribuidor) de seu produto aos canais de distribuição, os canais de distribuição, por sua vez, registram as entradas, o estoque, e o *sell out*, ou seja, as vendas do produto ao cliente final. Em outras palavras, a indústria, sem integração com os canais de distribuição, acompanha o seu produto apenas até o momento em que ele foi vendido aos seus distribuidores. No entanto, a informação de escoamento ao consumidor final do seu produto pode ser considerada estratégica para decisões que permeiam processos de manufatura, logística e comercial. Neste sentido, a integração dos dados do distribuidor pode permitir o acompanhamento de indicadores como *sell out* escoado, POG (Produto efetivamente aplicado), a rastreabilidade de lotes, SOC (*Share of Customer*), entre outros.

Um dos desafios identificados na integração de dados entre indústrias e seus canais de venda e distribuição é a normalização ou a correta associação dos produtos (HERRMANN et al., 2013).

Neste contexto, os atributos disponíveis em relação aos produtos como: ID do produto, código de barras, SKU (*Stock Keeping Unit* - Unidade de Manutenção de Estoque), nome do produto, descrição do produto, etc., são elementos fundamentais. Entretanto, devido a limitações de sistemas e características culturais de gestão, atributos que poderiam servir como chave unificada entre os agentes da cadeia, como o código de barras do produto, não são corretamente registrados ou ainda não possuem nenhum registro. Nestes casos, são necessárias análises dos atributos disponíveis e principalmente aqueles que são obrigatórios para a efetivação das movimentações fiscais com os produtos.

No Brasil, os atributos de identificação obrigatória de maior relevância semântica dos produtos são: a descrição do produto e a Nomenclatura Comum do Mercosul (NCM) (ENCAT, 2015), que fornece uma categorização de produto. A descrição do produto, por sua vez, se refere a um atributo livre, o qual é determinado por cada agente, ou seja, o mesmo produto pode possuir valores de descrições diferentes para CNPJs diferentes. A partir disso, torna-se necessário o emprego de soluções que permitam a realização de integração entre os dados de diferentes canais com relação aos dados da indústria. Esse processo é normalmente realizado através de análises manuais, feitas por especialistas de domínio, os quais, através de verificações e comparações individuais estabelecem relacionamentos entre as bases de dados de indústrias e canais de distribuição. Entretanto, soluções computacionais inteligentes podem ser exploradas como mecanismo de otimização e automação deste processo, como é o caso do emprego de métodos de análise de similaridade.

Neste artigo é proposta uma solução para a inferência de relacionamentos entre bases de dados distintas a partir da correlação entre atributos textuais. Para tanto, é apresentada uma proposta baseada no uso do método de distância de Levenshtein e uso da metodologia KDD. O método é avaliado em bases de dados reais de indústria e canais de distribuição do setor de agronegócios. Os resultados alcançados demonstram efetividade na identificação de relacionamentos acima de 80% dos registros.

2 Revisão bibliográfica

Conforme abordado por Bilenko, Basil e Sahami (2005), o uso de soluções computacionais voltadas para a correlação entre atributos de diferentes bases de dados é explorado na literatura científica sob diferentes nomes, tais como ligação de registros (FELLEGI; SUNTER, 1969), problema de *merge/purge* (WINKLER, 1999), detecção de duplicidades (MONGE; ELKAN, 1997; SARAWAGI; BHAMIDIPATY, 2002; BILENKO; MOONEY, 2003), correspondência de referência/citação (MCCALLUM; NIGAM; UNGAR, 2000; LAWRENCE; BOLLACKER; GILES, 1999), correspondência de nome de entidades e agrupamento (COHEN; RICHMAN, 2002), *hardening soft databases* (COHEN; KAUTZ; MCALLESTER, 2000), incerteza de identidades (PASULA et al., 2003), e leitura robusta (LI; MODE; ROTH, 2004).

Bilenko, Basil e Sahami (2005) destaca que a maior parte das soluções voltadas à integração ou correspondência de registros e diferentes bases de dados é tratada como um problema modular, consistindo de uma série de etapas a serem realizadas. Neste sentido, um dos principais procedimentos a serem realizados se refere ao cálculo da similaridade entre pares de registros que sejam potencialmente relacionados, onde podem ser utilizadas funções de cálculo de

distância entre termos, tais como o método de Levenshtein, utilizado como base no presente artigo, ou mesmo soluções baseadas em deep *learning*, por exemplo, como o método conhecido como Word2Vec (MIKOLOV et al., 2013). É importante destacar que o processo de correlação entre os diferentes atributos pode ser apoiado de forma manual, uma vez que grandes volumes de dados acabam por apresentar especificidades que demandam análises orientadas para sua identificação.

Neste contexto, a análise de similaridade entre termos pode ser considerada como um campo que desperta múltiplas oportunidades a serem exploradas. Em Kannan et al. (2011) foram aplicados métodos de similaridade para associação entre nomes desestruturados de produtos e seus nomes estruturados equivalentes dentro de uma plataforma de busca em *e-commerce*. A partir dos resultados, os autores realizaram análises descritivas em relação ao comportamento de busca de usuários mais precisas, e ainda, aprimoraram os algoritmos de recuperação de informações para que fossem gerados resultados de busca de maior qualidade. Dessa maneira, o usuário teria como acesso a as ofertas do produto certo mesmo utilizando uma vasta variedade de detalhes sobre ou seu nome, ou ainda se o mesmo deixasse de destacar algum termo importante do nome do produto.

Seguindo a mesma linha, em Dhana Lakshmi, Ramani e Eswara Reddy (2019) é apresentada uma abordagem baseada em aprendizado de máquina para o cálculo de similaridade entre elementos textuais referentes à recomendação de produtos em compras pela internet, no qual são combinadas diferentes funções lineares com o propósito de se otimizar os resultados alcançados. Neste trabalho os autores exploram tanto a descrição dos produtos quanto outros atributos como classificações provenientes dos consumidores. Já no domínio da saúde, por exemplo, métodos de cálculo de similaridade textual de registros são utilizados para identificação de semelhante entre anotações médicas, onde a partir da similaridade semântica entre trechos textuais, informações redundantes podem ser eliminadas, de modo a proporcionar uma diminuição da carga cognitiva em processos de tomada de decisão (WANG et al., 2018).

3 Metodologia

O principal objetivo deste artigo é avaliar o uso do método de distância de Levenshtein como solução para problemas de classificação de produtos a partir da similaridade entre seus nomes. Para tanto, a metodologia do trabalho será baseada no processo KDD (do inglês, *Knowledge Discovery in Databases*). Proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD consiste em um modelo para geração de informação aplicável a grandes volumes de dados que se baseia principalmente no uso de soluções computacionais inteligentes, especificamente no que se refere à métodos de mineração e dados e inteligência artificial. Neste sentido, o KDD possui a vantagem de ser um método iterativo e interativo, isto é, permite que intervenções sejam realizadas durante a execução de suas atividades, assim como o retorno à atividades anteriores (PROVOST; FAWCETT, 2016). As etapas do KDD são: seleção, pré-processamento, transformação, mineração, e interpretação de dados.

A fase de seleção refere-se à definição do problema a ser trabalhado e a partir disso a delimitação do conjunto de dados que deverá ser considerado para os propósitos de geração de informação. Neste sentido, considerando os aspectos que permeiam o estudo realizado neste tra-

balho, deverão ser considerados os atributos descritivos dos produtos tanto no âmbito da indústria quanto em seus canais de distribuição, assim como os valores de NCM relativos a cada um dos registros constantes da base. Para tanto, foi obtida uma base de dados contendo registros de indústrias do setor de agronegócios, assim como as bases de canais de distribuição. O nome da indústria e de seus canais de distribuição serão preservados devido à questões de proteção de propriedade industrial. A Figura 1 apresenta o diagrama entidade-relacionamento para a base de dados considerada. É interessante observar que não existe um atributo que permita o relacionamento direto entre as entidades, sendo este o problema central a ser resolvido através da proposta apresentada neste artigo.

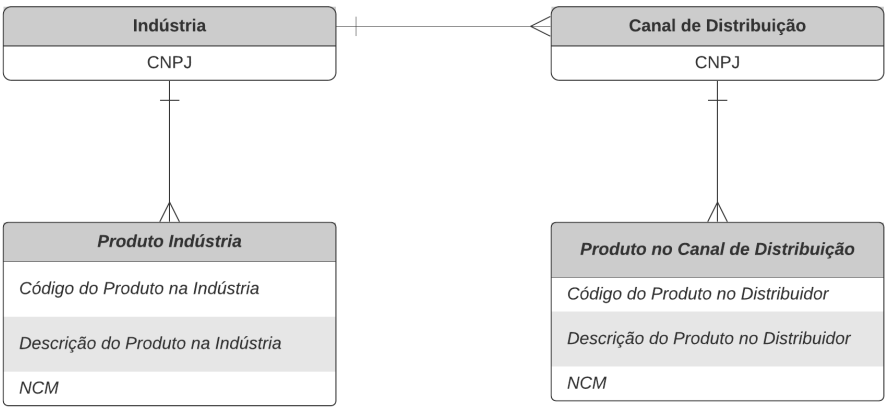

Figura 1: Diagrama Entidade-Relacionamento.
Fonte: autores.

No que se refere ao pré-processamento de dados, o primeiro passo foi realizar a identificação das ocorrências únicas em face do atributo descrição, considerando cada uma das instâncias de entidade indústria e entidade de distribuição. Essa etapa é importante devido ao fato de que na base de dados original existem uma série de registros que se referem ao mesmo produto, o que geraria processamento não necessário para os propósitos de classificação. Além da identificação dos registros únicos, durante a fase de tratamento foram removidos dados de quantificação dos produtos que eventualmente foram registrados junto ao atributo de descrição, tais como: medidas de peso ou volume, assim como *stop words*, isto é, termos não representativos para o processamento textual. Para isso, foi utilizada uma estratégia baseada no uso de expressões regulares, permitindo tanto a identificação de sequências de acordo com os padrões de unidades e medidas ocorrentes na base, como das *stop words*.

Os dados foram extraídos originalmente de um base de dados relacional através de um conjunto de consultas SQL e convertidos para o formato estruturado CSV (*comma separated values*). Essa transformação no formato de dados se justifica pela facilidade de manipulação frente às rotinas de programação utilizadas na solução, as quais se baseiam no uso da linguagem de programação Python, principalmente na biblioteca Pandas[1] e Numpy[2]. Uma das principais vantagens no uso de arquivos em formato CSV se refere ao fato de que estes apresentam os dados

[1] https://pandas.pydata.org/
[2] https://numpy.org/

em um formato textual aberto, promovendo uma flexibilidade e portabilidade em termos de uso de diferentes tipos de soluções computacionais (FERNANDES; CORDEIRO, 2016).

Para a realização da identificação de similaridades entre os dados de descrição do produto na indústria e no canal de distribuição, foi utilizado como base o método conhecido como Distância de Levenshtein. A Distância de Levenshtein é uma medida de comparação da similaridade entre elementos textuais dada pela quantidade de mudanças necessárias em um elemento para que este seja, ou se torne, igual ao outro. Esta medida foi originalmente utilizada para a recuperação de informações perdidas em transmissões binárias, onde se ocorriam inconsistências, inserções ou omissões, de sinais impactando na qualidade da informação recebida (LEVENSHTEIN, 1966). Com o avanço das tecnologias e a evolução de algoritmos de mineração de dados, a distância de Levenshtein continua se apresentando como uma solução arrojada para aplicação em problemas de similaridade. Dentre eles, classificação de texto e recuperação de expressões (LOPES, 2019), identificação de chaves entre diferentes fontes de dados (DORAN; WAMELEN, 2010), correção de erros, reconhecimento e comparação de padrões (SCHEPENS; DIJKSTRA; GROOTJEN, 2012; BEERNAERTS et al., 2019), entre outros.

O cálculo da Distância de Levenshtein aplicado a duas strings a e b, respectivamente de tamanhos $|a|$ e $|b|$, é definido pela função $\text{lev}_{a,b}(|a|,|b|)$, onde:

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j), & \text{se } \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j)+1 \\ \text{lev}_{a,b}(i,j-1)+1 \\ \text{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases}, & \text{caso contrário.} \end{cases} \tag{1}$$

onde $1_{(a_i \neq b_j)}$ é a função característica, a qual é igual a 0 sempre que $a_i = b_j$ e igual a 1 em caso contrário, e $\text{lev}_{a,b}(i,j)$ se refere à distância entre os primeiros i caracteres de a e os primeiros j caracteres de b. Na Tabela 1 é possível vislumbrar um exemplo de dois elementos, onde para o elemento 2 se tornar idêntico ao elemento 1 seria necessário a adição dos três caracteres ”ply”, ou seja, existe uma distância de três mudanças necessárias entre os elementos, sendo assim, o valor obtido através da distância de Levenshtein entre o elemento 1 e o elemento 2 é igual a três.

Elemento 1	**Elemento 2**	**Distância de Levenshtein**
Supply Chain	Super Chain	3

Tabela 1: Exemplo de aplicação do cálculo da Distância de Levensthein.

Para aplicação do método de Distância de Levenshtein para o cálculo da similaridade entre os atributos de descrição de produto de indústria e de canal de distribuição é necessário o emprego de procedimentos específicos ao conjunto de dados. Neste sentido, embora não exista uma correlação prévia, uma das características entre as instâncias dos atributos de descrição, em ambos contextos, é o fato de que estes sempre possuem o mesmo caractere inicial, o que pode ser utilizado como referência para refinamento nas rotinas de cálculo de similaridade, assim como ocorre com a equivalência do código NCM. Além disso, foi considerado um valor de *threshold* para a distância de Levenshtein calculada, ou seja, um valor máximo de distância com referência aos alvos candidatos. O Algoritmo 1 apresenta, em detalhes, os procedimentos aplica-

dos para resolução do problema de cálculo de similaridade para integração entre dados empresa e canais de distribuição.

Algoritmo 1: Similaridade textual baseada em Distância de Levenshtein.

```
   Dados: Base de produtos da indústria; Base de produtos dos canais de distribuição.
   Resultado: Base com correlação entre produtos da indústria e produtos dos canais de
              distribuição.
 1 Realizar rotinas de pré-processamento e transformação de dados;
 2 para cada produto do distribuidor faça
 3     para cada produto da indústria faça
 4         se NCM produto da indústria == NCM produto do distribuidor então
 5             Verificar se primeiros caracteres são equivalentes;
 6             se equivalentes então
 7                 Calcular Distância de Levenshtein;
 8             fim
 9         fim
10     fim
11     se existe produto(s) com Distância de Levenshtein menor que o treshold então
12         Associar produto do distribuidor com menor distância ao produto da indústria;
13     senão
14         Associar produto do distribuidor com "não encontrado";
15     fim
16 fim
```

4 Resultados e discussões

A partir dos aspectos metodológicos determinados, foi obtida uma base de dados contendo alvos de 84 indústrias, com um total de 5.216 registros. Os alvos estão segmentados em seis diferentes classes, identificadas e quantificadas conforme demonstrado na Tabela 2. Apesar de não haver um balanceamento em termos da quantidade de elementos por classe, considerando as características da solução proposta, a qual não se refere a um método em que seja necessário realizar treinamento a partir de dados, não há influências nos resultados obtidos.

Segmento	Quantidade
D1	65
D2	511
D3	607
D4	352
D5	71
S1	3610

Tabela 2: Dados da indústria.
Fonte: dados de pesquisa.

Os dados referentes aos canais de distribuição foram obtidos a partir de 1.002 CNPJs diferentes, apresentando um total de 241.429 registros de produtos, os quais se referem a 2.335 NCMs.

Apesar da disponibilidade de todo o universo de dados, para fins de inferência dos possíveis relacionamentos entre as entidades e produtos da indústria com os canais de distribuição não é necessário que as rotinas de cálculo de similaridade sejam aplicadas a todos estes, uma vez que existe a possibilidade de duplicidade nos dados provenientes dos canais de distribuição, ou seja, um produto pode ser registrado mais de uma vez em uma mesma venda e em diferentes vendas. Além disso, é necessário ainda verificar a correspondência dos registros em relação aos NCMs pertencentes às famílias consideradas no experimentos, família D (segmentos D1, D2, D3, D4 e D5) e família S (segmento S1), sendo que os registros que não pertencem a estes NCMs foram desconsiderados da base. Considerando isso, após a etapa de pré-processamento a base de dados de produtos dos canais de distribuição resultou em um total de 34.928 registros, sendo 29.184 pertencentes aos segmentos da família D, e 5.744 à família S.

De acordo com os procedimentos apresentados no Algoritmo 1, os dados pré-processados e transformados foram submetidos às rotinas de inferência das relações entre os registros das entidades de produto na indústria e nos canais de distribuição. Neste sentido, para os registros pertencentes à família D, o método proposto possibilitou a identificação efetiva de 23.876 instâncias distintas, o que representa 84,8% dos dados. Para a família S, a solução possibilitou a identificação de 4.766 instâncias, ou seja, 82,9% da base de dados. Os registros que não obtiveram inferência, foram identificados e apresentados em relatório para a realização de análises posteriores.

5 Conclusão

O presente artigo apresentou uma solução para a identificação de relacionamentos entre entidades distintas de produtos, com base no uso de cálculo de similaridade textual utilizando o método de distância de Levenshtein. O método foi avaliado através de aplicação em registros provenientes de bases de dados reais, obtidas a partir de indústria do setor do agronegócio, assim como canais de distribuição associados a esta. Os resultados alcançados demonstraram a efetividade do método, o qual conseguiu, para ambas as famílias aplicadas, obter uma identificação maior que 80%. Apesar de não terem sido feitos cálculos de acurácia, os relacionamentos identificados foram apresentados à indústria e aos canais de distribuição, e não foram realizadas contestações sobre os resultados por parte destas.

Além disso, o relacionamento realizado entre as visões de produto da indústria e de seus canais de distribuição proporcionou um melhor acompanhamento das movimentações contidas no fluxo da cadeia de suprimentos, possibilitando uma visualização mais clara e completa para indústria em relação ao comportamento dos seus produtos considerando os períodos de estoque e escoamento dos mesmos dentro de seus distribuidores. O que, consequentemente, apontou para a efetividade da integração de dados entre diferentes agentes da cadeia de suprimento, como uma ferramenta aliada à Gestão da Cadeia de Suprimentos a qual pode agregar valor aos processos de manufatura e desenvolvimento de serviços.

Futuramente, se pretende realizar a avaliação dos resultados através do cálculo da acurácia e do método de matriz de confusão. Além disso, se pretende ainda comparar os resultados com outros métodos de cálculo de similaridade textual, assim como aprimorar a etapa de pré-processamento de dados, de modo a se levar em conta aspectos e padrões que foram obtidos

através de análises dos registros aos quais não foram obtidas correspondências, nos quais se identificou um considerável uso de siglas para representação de descrições dos produtos.

Referências

BALLOU, Ronald H. **Gerenciamento da Cadeia de Suprimento/Logística Empresarial**. Porto Alegre: Bookman, 2007.

BEERNAERTS, Jasper et al. A method based on the Levenshtein distance metric for the comparison of multiple movement patterns described by matrix sequences of different length. **Expert Systems With Applications**, v. 115, p. 373–385, 2019.

BILENKO, M.; BASIL, S.; SAHAMI, M. Adaptive product normalization: using online learning for record linkage in comparison shopping. In: FIFTH IEEE International Conference on Data Mining (ICDM'05). [S.l.: s.n.], 2005.

BILENKO, M.; MOONEY, R. J. Adaptive duplicate detection using learnable string similarity measures. In: PROC. of ACM SIGKDD 2003. [S.l.: s.n.], 2003. p. 39–48.

BOWERSOX, Donald J. et al. **Gestão Logística da Cadeia de Suprimentos**. Porto Alegre: AMGH, 2014.

COHEN, W. W.; KAUTZ, H.; MCALLESTER, D. Hardening soft information sources. In: PROC. of ACM SIGKDD 2000. [S.l.: s.n.], 2000. p. 255–259.

COHEN, W. W.; RICHMAN, J. Learning to match and cluster large high-dimensional data sets for data integration. In: PROC. of ACM SIGKDD 2002. [S.l.: s.n.], 2002. p. 475–480.

DHANA LAKSHMI, P.; RAMANI, K.; ESWARA REDDY, B. Neighborhood Algorithm for Product Recommendation. In: COMPUTATIONAL Intelligence and Big Data Analytics: Applications in Bioinformatics. Singapore: Springer Singapore, 2019. p. 37–53.

DORAN, Harold C.; WAMELEN, Paul B. Van. Application of the Levenshtein Distance Metric for the Construction of Longitudinal Data Files. **Educational Measurement: Issues and Practice**, v. 29, n. 2, p. 13–23, 2010.

ENCAT. **Orientação de Preenchimento da NF-e - versão 2.02**. [S.l.], 2015.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. **Journal of the American Statistical Association**, v. 64, n. 328, p. 1183–1210, 1969.

FERNANDES, J. L. F.; CORDEIRO, D. F. Avaliação de formatos de publicação de dados abertos governamentais através de indicadores de usabilidade. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, n. 1, p. 65–84, 2016.

GOMES, Carlos Francisco Simões; RIBEIRO, Priscilla Cristina Cabral. **Gestão da Cadeia de Suprimentos Integrada à Tecnologia da Informação**. São Paulo: Cengage Learning, 2013.

HERRMANN, Felipe Fehlberg et al. Benefícios e impeditivos à integração da cadeia de suprimentos calçadista por meio da tecnologia de informação. **Gestão & Produção**, v. 20, n. 4, p. 939–952, 2013.

KANNAN, Anitha et al. Matching Unstructured Product Offers to Structured Product Specifications. In: PROCEEDINGS of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.: s.n.], 2011.

LAWRENCE, S.; BOLLACKER, K.; GILES, C. L. Autonomous citation matching. In: PROC. of Agents-1999. [S.l.: s.n.], 1999. p. 392–393.

LEVENSHTEIN, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics Doklady**, v. 8, n. 4, p. 707–710, 1966.

LI, X.; MODE, P.; ROTH, D. Robust reading: Identification and tracing of ambiguous names. In: PROC. of NAACL-2004. [S.l.: s.n.], 2004. p. 17–24.
LOPES, Inês Margarida Silva Paz. **Qualidade dos dados & Machine Learning : uma nova abordagem aos censos populacionais e habitacionais**. 2019. f. 40. Dissertação (Mestrado em Gestão de Sistemas de Informação) – Universidade de Lisboa, Instituto Superior de Economia e Gestão.
MCCALLUM, A.; NIGAM, K.; UNGAR, L. Efficient clustering of high-dimensional data sets with application to reference matching. In: PROC. of ACM SIGKDD-2000. [S.l.: s.n.], 2000. p. 169–178.
MIKOLOV, Tomas et al. Efficient Estimation of Word Representations in Vector Space. In: PROCEEDINGS of the International Conference on Learning Representations (ICLR 2013). [S.l.: s.n.], 2013.
MONGE, A. E.; ELKAN, C. P. An efficient domainindependent algorithm for detecting approximately duplicate database records. In: PROC. of SIGMOD 1997 Workshop on Research Issues on Data Mining and Knowledge Discovery. [S.l.: s.n.], 1997. p. 23–29.
PASULA, H. et al. Identity uncertainty and citation matching. In: NIPS 15. [S.l.: s.n.], 2003. p. 1401–1408.
PROVOST, F.; FAWCETT, T. **Data Science para Negócios**. Rio de Janeiro: Alta Books, 2016.
SARAWAGI, S.; BHAMIDIPATY, A. Interactive deduplication using active learning. In: PROC. ACM SIGKDD-2002. [S.l.: s.n.], 2002. p. 269–278.
SCHEPENS, Job; DIJKSTRA, Ton; GROOTJEN, Franc. Distributions of cognates in Europe as based on Levenshtein distance. **Bilingualism: Language and Cognition**, v. 34, n. 3, p. 529–532, 2012.
SÊMOLA, Marcos. **Gestão da Segurança da Informação: uma visão executiva**. Rio de Janeiro: Elsevier, 2014.
SIMCHI-LEVI, David; KAMINSKY, Philip; SIMCHI-LEVI, Edith. **Cadeia de Suprimentos Projeto e Gestão**. Porto Alegre: Bookman, 2010.
WANG, Yanshan et al. MedSTS: a resource for clinical semantic textual similarity. **Language Resources and Evaluation**, v. 54, p. 57–72, 2018.
WINKLER, W. E. **The state of record linkage and current research problems**. Washington, DC, 1999.