Um Estudo Sobre a Falta de Padronização na Descrição de Produtos em Notas Fiscais Eletrônicas

João D. R. Mazzarolo[1], **Raul Steinmetz**[1], **Sergio L. S. Mergen** [2]

[1]PET-CC – Centro de Tecnologia – Universidade Federal de Santa Maria (UFSM)
97105-900 – Santa Maria – RS – Brazil

[2]Departamento de Linguagens e Sistemas de Computação
Universidade Federal de Santa Maria (UFSM)

{rsteinmetz, jdmazzarol, mergen}@inf.ufsm.br

Abstract. *Due to the institutionalized computerization of commercial systems, information regarding products purchased by users are increasingly being made available in digital format, through electronic invoices. With access to this data, it is possible to build different types of services, such as purchasing portals, financial coaching tools and big data analytics. However, there is no standardization in the definition of the names that should be used to represent the products. This lack of standardization makes it difficult for integration processes to consolidate products sold by different establishments. This article aims to investigate how this problem occurs in real scenarios, specifically considering products sold by the supermarket sector. To achieve this end, existing products from a collection of invoices extracted from the Nota Fiscal Gaúcha project were analyzed. The analysis presents statistics regarding the most commonly encountered problems and highlights textual similarity techniques that can help clean up/recognize ill-behaved data.*

Resumo. *Com a informatização institucionalizada de sistemas comerciais, informações referentes a produtos adquiridos por usuários estão sendo cada vez mais disponibilizadas em formato digital, através de notas fiscais eletrônicas. Com o acesso a esses dados é possível construir diversos tipos de serviço, como portais de compra, ferramentas de coaching financeiro e análises de big data. No entanto, não existe padronização na definição dos nomes que devem ser usados para representar os produtos. Essa falta de padronização traz dificuldades para processos de integração que desejam consolidar produtos vendidos por estabelecimentos diferentes. Este artigo tem como propósito investigar como esse problema se manifesta em cenários reais, considerando especificamente produtos vendidos pelo setor supermercadista. Para isso, foram analisados produtos existentes em uma coleção de notas fiscais coletadas a partir do programa Nota Fiscal Gaúcha. A análise apresenta estatísticas referentes aos problemas mais comumente encontrados e destaca técnicas de similaridade textual que podem ajudar a fazer a limpeza/reconhecimento dos dados disformes.*

1. Introdução

A nota fiscal é um documento obrigatório de existência digital, gerado e armazenado eletronicamente pela Receita Federal do Brasil, pelas prefeituras ou por outras entidades

conveniadas, para registrar as operações comerciais e de prestação de serviços. Com o passar do tempo, as notas fiscais, originalmente disponibilizadas em papel, vem sendo substituídas por notas fiscais eletrônicas (NFS-e's). Dentre os impactos positivos dessa mudança, pode-se citar a facilidade da obtenção de uma nota formatada e a possibilidade de utilização de algoritmos que extraiam esses dados para usá-los na criação de diversos tipos de serviço, como portais de compra, ferramentas de coaching financeiro e análises de big data.

Na lei nº 4.502 de 30 de Novembro de 1964 está presente o seguinte artigo, que estabelece regulações ligadas ao formato das notas fiscais: "Art . 48. A nota fiscal obedecerá ao modêlo que o regulamento estabelecer e conterá as seguintes indicações mínimas: I - denominação "Nota Fiscal"e número de ordem; II - nome, endereço e número de inscrição do emitente; III - natureza da operação; IV - nome e enderêço do destinatário; V - data e via da nota e data da saída do produto do estabelecimento emitente; VI - discriminação dos produto pela quantidade, marca, tipo, modêlo, número, espécie, qualidade e demais elementos que permitam a sua perfeita identificação, assim como o preço unitário e total da operação, e o preço de venda no varejo quando o cálculo do impôsto estiver ligado a êste ou dêle decorrer isenção; VII - classificação fiscal do produto e valor do impôsto sôbre êle incidente; VIII - nome e enderêço do transportador e forma de acondicionamento do produto (marca, numeração, quantidade, espécie e pêso dos volumes)[sic]".

Por mais que haja uma regulamentação clara no tocante à formatação de notas, os documentos fiscais emitidos no Brasil apresentam um grande problema: a falta de padronização na descrição de produtos. Esta é uma questão notável, especialmente no setor supermercadista, onde produtos similares ou até mesmo iguais têm nomes muito distintos, seja pela mudança de supermercado fornecedor ou pela mudança da marca. Tal disparidade traz dificuldades para processos de integração que desejam consolidar produtos vendidos por supermercados diferentes.

Neste contexto, o objetivo do artigo é analisar os problemas existentes na identificação de produtos em notas fiscais eletrônicas. Para atingir esse fim foi empregado um processo de coleta, extração e análise de produtos existentes em mais de uma centena de notas fiscais obtidas a partir do programa Nota Fiscal Gaúcha. Neste trabalho focou-se em produtos oferecidos pelo setor supermercadista.

O artigo está estruturado da seguinte forma: a seção 2 apresenta trabalhos relacionados. A seção 3 descreve a proposta como um todo. A seção 4 apresenta os experimentos, onde destacam-se as diferenças encontradas em produtos similares vendidos em supermercados, e o desempenho de funções comumente usadas para descobrir similaridade entre strings. Por fim, a seção 5 traz as conclusões alcançadas e realça possibilidades de trabalhos futuros.

2. Trabalhos Relacionados

Vários autores já apresentaram pesquisas envolvendo notas fiscais, sua importância, vantagens e desvantagens na economia moderna. Nesse âmbito há o trabalho de [Vieira et al. 2019], que visa verificar se a implantação do programa de Nota Fiscal Eletrônica (NFS-e) gerou como consequência algum incremento na arrecadação do Estado de Goiás utilizando testes de médias e estimação de regressões *difference-in-differences*. Há também o trabalho de [Neto and Martinez 2016], que faz uma análise dos dados de

arrecadação tributária nas maiores cidades do brasil com a finalidade de avaliar se houve um aumento na arrecadação.

Além da bibliografia em torno de notas fiscais, alguns autores já discutiram métodos de integração de pseudo-padrões de descrição de produtos, com foco no e-commerce. À exemplo de trabalho tem-se [Zhang 2014], que mostra o problema da falta de padronização de descrições em diferentes comércios e a necessidade atual de ferramentas que permitam a comunicação de dados em modelos *Business to Business*. O trabalho de [Fensel et al. 2001] também foca nas dificuldades relacionadas à integração de descrições de produtos de diferentes fontes. No artigo é explicado o processo da extração e organização de informações e as brechas que ainda persistem, como falta de informações nas descrições não padronizadas de produtos e a possibilidade de uso de Inteligência Artificial para identificar informações falsas.

Nesse mesmo sentido, mas com foco na solução do problema em questão tem-se o trabalho de [Bergamaschi et al. 2002], que discute o fato de que a existência de várias propostas de padrões de descrição causa descrições cada vez mais individuais e peculiares de produtos. A proposta principal do trabalho é a apresentação e sugestão de uso de uma metodologia semi-automática que consegue integrar e mapear diferentes padrões de classificação, possibilitando a integração de dados vindos de fontes heterogêneas. O método demonstrou resultados interessantes mapeando entre produtos classificados por dois padrões distintos (UNSPSC - Electronic Commerce Code Management Association e ECLASS – Standard for Mater Dara and Semantics for Digitalization) usados por cadeias de fornecimento de produtos.

Até onde sabemos, por mais que existam aplicativos que usam algoritmos de comparação de produtos, como a própria Nota Fiscal Gaúcha, assim como artigos que apresentam ideias e métodos de integração de produtos encontrados no âmbito do *e-commerce*, não existem trabalhos relacionados com a problematização ou solução da disformidade na descrição dos nomes de produtos vendidos pelo setor supermercadista. Tal fato suscita a relevância do artigo proposto.

3. Método Proposto

O propósito deste trabalho é explorar as diferenças presentes na descrição dos produtos nas notas fiscais de diferentes estabelecimentos. Para atingir este objetivo fazem-se necessárias a coleta de notas fiscais diversas e a extração de seus dados, além da análise das informações.

A coleta de notas fiscais utilizou o site Nota Fiscal Gaúcha (Sefaz RS), programa que tem como objetivo incentivar os cidadãos do Rio Grande do Sul a exigirem a inclusão do CPF na emissão de notas fiscais. Esse site disponibiliza notas fiscais eletrônicas formatadas e em modelo PDF geradas quando o cidadão cadastra seu CPF na nota fiscal de um estabelecimento. A partir da coleta, parte-se para as demais etapas, descritas na Figura 1: (a) Extração do conteúdo; (b) Rotulagem dos produtos; e (c) Análise dos dados. Cada um destas etapas será apresentada a seguir.

3.1. Extração de conteúdo

A extração do conteúdo das notas foi realizada computacionalmente através de um algoritmo *Parser*, que analisa uma sequência (texto da NFS-e, nesse caso) e realiza uma

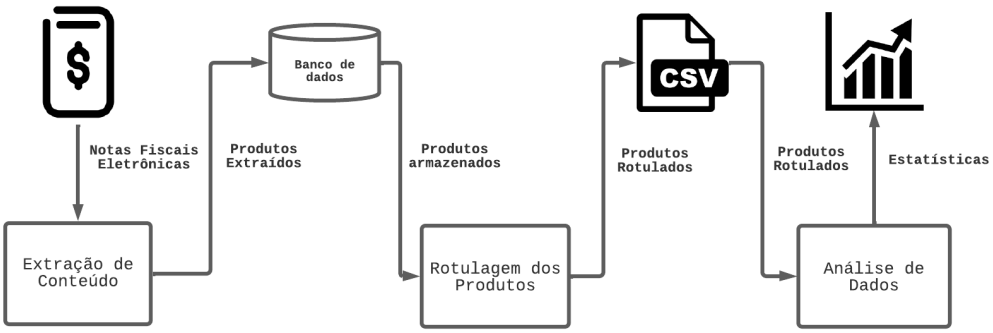

Figura 1. Diagrama do Método Desenvolvido

análise gramatical segundo uma determinada gramática. Os dados reconhecidos pelo *parser* são extraídos e armazenados em um banco de dados.

Com relação ao código de parsing foi utilizada a linguagem de programação Java, com auxílio da biblioteca PDFBox (biblioteca de código aberto ultilizada para criar, renderizar, imprimir, dividir, mesclar, alterar, verificar e extrair texto e metadados de arquivos em formato PDF).

O banco de dados foi feito na linguagem MYSQL. Os dados guardados contém informações essenciais para a análise, como nome do produto, estabelecimento fornecedor da nota fiscal, preço e unidade de medida (quilograma ou unidade).

3.2. Rotulagem dos Dados

Um dos principais objetivos deste trabalho é identificar, para cada produto, todas as ocorrências que aparecem em notas fiscais com divergência na sua representação textual. Para isso é necessário que os produtos sejam rotulados. Ou seja, para cada produto deve-se estabelecer um nome canonizado que sirva como identificador único daquele item. Neste trabalho, este nome representativo será chamado de produto de referência.

Afim de facilitar a etapa da rotulagem foi utilizado um algoritmo de clusterização que faz um agrupamento inicial de produtos que possam ter alguma relação. O algoritmo funciona da seguinte forma: para cada produto p_i não clusterizado cria-se um novo cluster que irá conter p_i e todos os produtos não clusterizados p_j que tenham correspondência com p_i. Para que haja correspondência, os produtos p_j são divididos em termos, usando o caractere de espaço como delimitador. Caso o primeiro termo do produto p_i apareça como substring de algum termo de outro produto, uma correspondência é encontrada. Por exemplo, para o produto p_i "BATATA BRANCA GRANEL", o primeiro termo é "BATATA". Alguns produtos para os quais o algoritmo encontraria correspondência são "BATATA ROSA"e "BATATA BRANCA".

Alguns dos grupos formados por esse algoritmo são: (a)"PEPSI TRADICIONAL 600ML", "PEPSI TRADICIONAL 2L", "PEPSI ZERO 2L"(b) "BANANA CATURRA GRANEL", "BANANA BRANCA KG", "BALA PIETROBON 250G BANANA", "BANANA PRATA", "BANANA PRATA GRANEL'", (c) "QUEIJO MUSSARELA PRESIDENT AT", "QUEIJO PARMESAO S", "PALITO QUEIJO 220G KROCKS", "QUEIJO SCALA", "QUEIJO COLONIAL P", "QUEIJO".

Como pode-se ver, alguns grupos formados pela clusterização automática contém

elementos que ficariam melhor representados em outro grupo (ex. "BALA PIETROBON 250G BANANA"). Por esse motivo, os grupos passaram por uma etapa adicional de clusterização manual visando criar grupos mais coesos, Durante esta etapa, uma das principais preocupações foi de agrupar produtos que, na visão do usuário, satisfaçam seu critério de busca em um sistema de informação. Por exemplo, caso um usuário busque por "BANANA PRATA", "BANANA PRATA KG"seria considerado um resultado relevante, enquanto "BANANA CATURRA KG"não, por representar um produto diferente daquele que foi buscado.

Levando em conta esta preocupação, como regra geral decidiu-se que marcas diferentes devem gerar agrupamentos diferentes (desde que a marca esteja devidamente evidenciada no nome do produto). Por exemplo, "TORRADA INTEGRAL BAU"e "TORRADA INTEGRAL ISA"seriam parte de grupos diferentes, por representarem produtos das marcas Bauducco e Isabela, respectivamente. Além disso, produtos com caracterizações bem especificas também são separados. Por exemplo, "PAO INTEGRAL"e "PAO DE MILHO MULT"ficariam em grupos separados.

Após a clusterização, foi realizado o processo manual de rotulagem dos dados. O processo consistiu em, para todos os produtos de um mesmo grupo, atribuir um produto de referência, que sirva como generalização de todos os produtos do grupo. Além da atribuição de um nome de referência, os produtos também foram categorizados de acordo com o tipo do produto. Ao todo, onze categorias foram utilizadas. São elas: farináceos, frutas, vegetais, bebidas, doces e salgados, sementes e grãos, frios e derivados de animais, não alimentícios, leite e derivados, prontos para consumo (geralmente encontrados em padarias) e auxiliares de cozinha.

A Tabela 1 apresenta alguns produtos após a etapa de rotulagem. A tabela apresenta alguns dos produtos divididos em dois grupos ("PAO INTEGRAL"e "PAO DE MILHO MULT"), sendo que os dois grupos pertencem à categoria dos "FARINÁCEOS"). A listagem completa contendo categoria e produto de referência está disponível em CSV no site do projeto[1].

Produto	Referência	Categoria
PAO INTEGRAL SAN	PAO INTEGRAL	FARINÁCEOS
PAO INTEGRAL	PAO INTEGRAL	FARINÁCEOS
PAO INTEGRAL E SOVADO	PAO INTEGRAL	FARINÁCEOS
PAO MULTIGRAOS	PAO INTEGRAL	FARINÁCEOS
PAO LEV SEMI INTEG S	PAO INTEGRAL	FARINÁCEOS
PAO DE MILHO MULT	PAO DE MILHO MULT	FARINÁCEOS
PAO MILHO EMERY 5	PAO DE MILHO MULT	FARINÁCEOS

Tabela 1. Parte do Arquivo CSV

Convém destacar que as regras utilizadas para a rotulagem são subjetivas e não passaram por nenhuma avaliação empírica. Mesmo assim, elas serviram como base para os experimentos conduzidos e permitiram que algumas estatísticas e questões bastante peculiares pudessem ser levantadas, como será mostrado na Seção 4. Como trabalho futuro, pretende-se estudar melhor este problema, visto que a rotulagem de dados pode ser útil para diversos problemas de aprendizado de máquina.

[1] https://github.com/Mazzarolo/Analise-de-Produtos-em-Notas-Fiscais

3.3. Análise dos Dados

A partir do arquivo CSV gerado foram feitas diferentes análises computacionais utilizando a linguagem Python. As análises empregaram funções de similaridade textual que visam descobrir as semelhanças entre produtos que pertençam ao mesmo grupo. Duas funções foram usadas:

- **Edit Distance (ED)**: Retorna o número de operações de edição necessárias para transformar uma string em outra. O valor retornado é normalizado entre zero e um através do complemento entre a divisão do número de operações pelo tamanho da maior string. Um maior detalhamento sobre a função pode ser encontrado no trabalho de [Levenshtein et al. 1966].
- **Longest Common Subsequence (LCS)**: Retorna o número de caracteres não consecutivos e ordenados em comum entre duas strings. O valor retornado é normalizado entre zero e um através da divisão do número de caracteres em comum pelo tamanho da menor string. Um maior detalhamento sobre o método pode ser encontrado no trabalho de [Bergroth et al. 2000].

A função ED é mais indicada para detectar semelhanças em casos onde as diferenças entre as strings é sutil, como variações na concordância nominal ou erros ortográficos. Por exemplo, "PITAYA KG"e "PITAIA KG"são nomes usados em estabelecimentos comerciais diferentes, e cuja similaridade é detectada pela função ED.

Já a função LCS usa a menor string no denominador para detectar as semelhanças em casos onde uma string é uma subsequência da outra, como em abreviaturas e prefixos. Como exemplo do uso de prefixos tem-se as descrições "CHOCOLATE NESTLE 33G PRESTIGIO"e "CHOC NESTLE PRESTIGIO 33G", onde "CHOC"é um prefixo usado para representar a palavra "CHOCOLATE". Como exemplo do uso de abreviaturas, tem-se as descrições "QJ COL PREMIA KG"e "QUEIJO COLONIAL P", onde "QJ"é uma abreviatura para representa a palavra "QUEIJO".

4. Resultados

Esta seção apresenta análises referentes à diferenças na representação de nomes de produtos vendidos em estabelecimentos comerciais do setor supermercadista. Ao todo, foram coletadas 138 notas fiscais de supermercados localizados no Rio Grande do Sul e Santa Catarina, nas cidades de Caxias do Sul(RS), Santa Rosa(RS), Santa Maria(RS), Urubici(SC). Entre os estabelecimentos estão "Companhia Zaffari Comércio e Industria", "Timy Industria Comercio Alimentos Eireli"e "Carrefour Comércio e Industria LTDA", entre outros. A escolha de diversos estabelecimentos diferentes tem como objetivo encontrar uma maior diversificação de descrições para os mesmos produtos, enriquecendo a análise. As notas compreendem o período entre Janeiro de 2021 e Fevereiro de 2022.

Como mencionado na seção 3.2, os produtos passaram por uma etapa de rotulagem que definiu uma categoria geral e agrupou-os em torno de algum produto de referência. Para a análise, foram removidos grupos com apenas um produto. A Tabela 2 demonstra estatísticas a respeito dos dados que restaram após a remoção. A partir destes dados, foram selecionadas categorias que possuíssem mais do que 30 produtos e 10 grupos (indicadas em negrito). Essa medida foi tomada para que as análises fossem embasados em torno de volumes mais significativos de dados.

Categoria	Número de Produtos	Número de Grupos
AUXILIARES DE COZINHA	15	7
BEBIDAS	16	8
DOCES E SALGADOS	40	18
FARINÁCEOS	33	14
FRIOS E D.	50	21
FRUTAS	55	19
LEITE E D.	39	16
NÃO ALIMENTÍCIOS	38	15
PRONTOS	10	5
SEMENTES E GRÃOS	8	4
VEGETAIS	51	15

Tabela 2. Dados sobre a Amostragem após Filtragem

4.1. Número de representações distintas por produto

Neste experimento, o objetivo é analisar a quantidade de variações de descrição usadas para cada grupo. A Figura 2 apresenta os resultados alcançados.

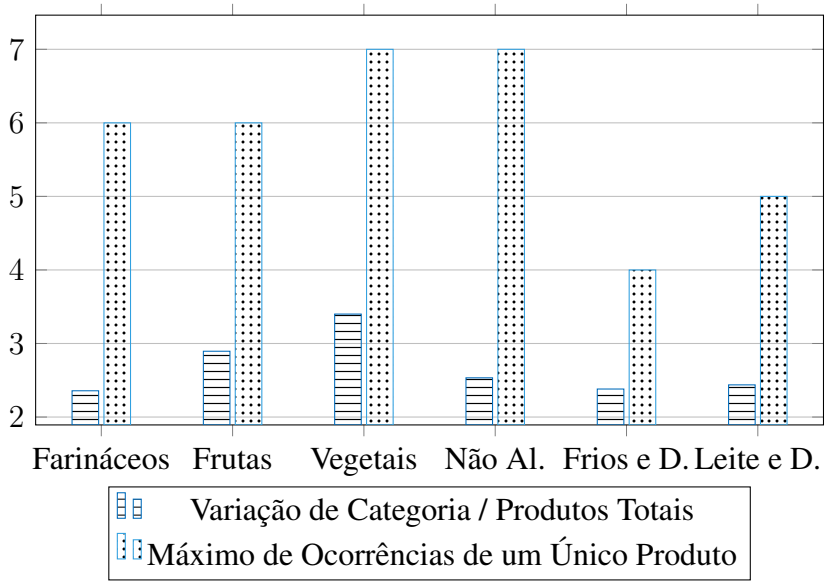

Figura 2. Quantidade de Variações de Nomenclatura Média e Máxima por Categoria

Para cada categoria, a quantidade de variações exibida refere-se à média de variações considerando todos os seus respectivos grupos de produtos. Já o máximo refere-se ao grupo da categoria que obteve a maior quantidade de variações. Por exemplo, na categoria 'Frutas', os grupos de produtos possuem na média três descrições distintas, enquanto que o grupo que obteve uma maior quantidade de variações foi o da "BANANA PRATA", com seis ocorrências ("BANANA PRATA", "BANANA PRATA GRANEL", "BANANA PRATA KG", "BANANA PRATA PCT 800", "BANANA PRATA COMUM C", "BANANA PRATA COMUM").

Como se pode ver, as categorias com maior média de variações são as frutas e os vegetais. A quantidade de notas coletadas é pequena demais para afirmar que esta é

uma tendência. Com um número maior de notas, possivelmente a quantidade de variações nas demais categorias poderia crescer a ponto de equiparar-se ou ultrapassar as frutas e vegetais. De qualquer forma, o experimento serviu para mostrar que a questão da variação existe em todas as categorias, e isso caracteriza um desafio que deve ser encarado em processos de integração.

4.2. Distância de edição entre produtos de um mesmo grupo

A seção anterior visou descobrir a quantidade média de variações na descrição de um produto. Nesta seção, o objetivo é verificar o quão disformes são estas variações. Para estudar esta característica foi usada a distância de edição. Para cada grupo foi encontrada a distância de edição entre o produto de referência do grupo e todas as suas variações. Quanto maior a distância entre as variações, mais disformes elas são. Por exemplo, no grupo da "BANANA PRATA", a distância média foi de 5,33.

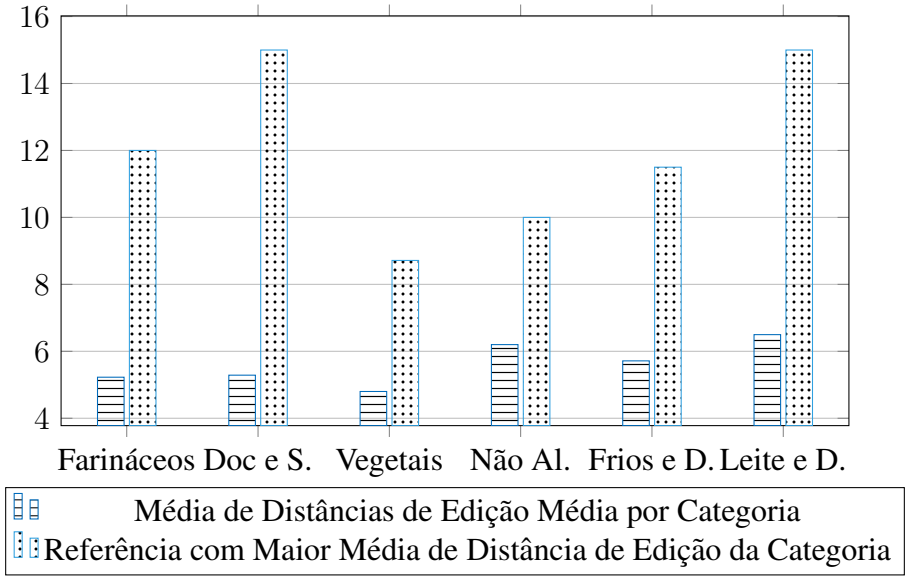

Figura 3. Distância de Edição de Produtos Média e Distância Máxima por Categoria

A Figura 3 apresenta os resultados alcançados, compilados por categoria. Para cada categoria, encontrou-se a distância média considerando as distâncias médias de todos os seus grupos. Por exemplo, na categoria dos vegetais, a distância média foi de 4,79. Já o valor máximo exibido na figura se refere ao produto da categoria que obteve a maior distância de edição em relação ao seu produto de referência.

Os grupos que se destacam com as maiores médias são: "Frios e Derivados de Animais"e "Leite e Derivados". Esse panorama revela que esses grupos possuem uma falta de padronização maior que os demais, porém, de uma forma não tão expressiva, visto que seus números não são tão maiores quando comparados ao restante.

Entretanto, ao observar os produtos com maior distância de edição de cada categoria pode-se perceber que estes valores são sempre muito maiores que a média de sua categoria, revelando a existência de nomenclaturas bastante disformes dentro de todas as categorias. Sob este viés, pode-se concluir que os casos com maior falta de padronização estão muito mais relacionados a produtos específicos do que a categorias.

Como exemplo, pode-se citar o produto "LING TOSCANA GRANBERG APERRESF", cuja distância de edição é igual a 11, em relação ao seu similar "LINGUICA SUINA GRANBERG RESF", enquanto a média de distância de edição da categoria leite de frios e derivados é 5,71. Outro aspecto interessante a respeito deste caso específico é a ocorrência de substrings em comum (("LINGUICA", "LING"], ["RESF", "APERRESF"]). Funções que valorizem substrings, como a LCS, poderiam ser usadas para aumentar a similaridade, em casos como este.

4.3. Precisão de funções de cálculo de similaridade

A seção anterior exibiu um caso em que a função de similaridade LCS seria mais apta a encontrar a similaridade entre variações de nome do que a função ED. Nesta seção, o intuito é verificar o desempenho destas duas funções quando usadas para realizar a recuperação de produtos com base em busca por palavra-chave.

Afim de gerar os resultados, executou-se uma série de consultas distintas sobre a base de produtos coletados. Cada produto coletado foi usado como palavra-chave em uma consulta, e se analisou os resultados retornados dentro do top-k (ranking de k produtos com maior semelhança sintática à referência), com k variando de um a cinco. A partir das respostas calculou-se a taxa de acerto, que indica quantas das k respostas realmente se referem a variações do produto usado como palavra-chave. Quanto maior a taxa de acerto, mais precisa é a função de busca.

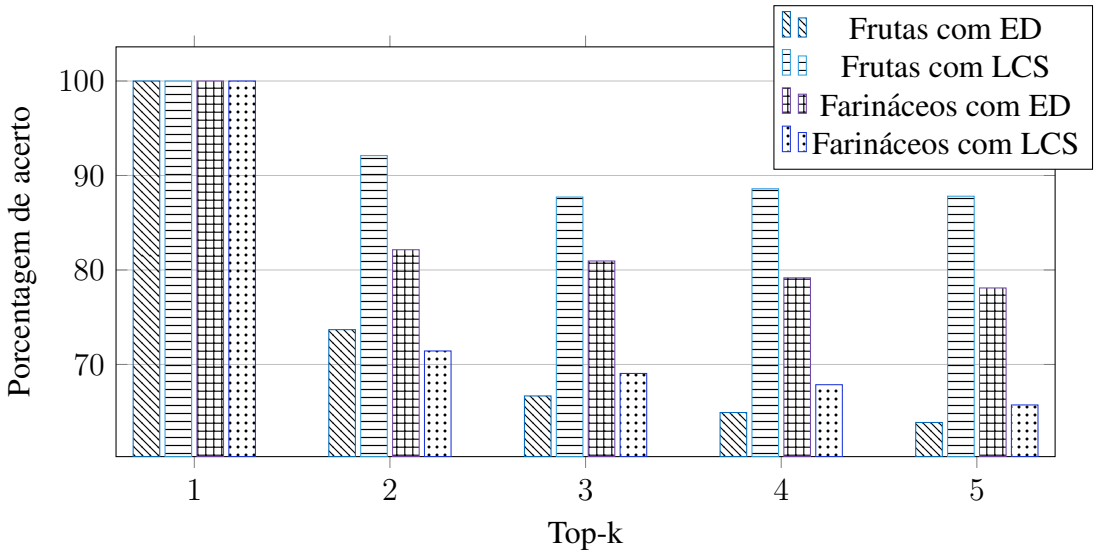

Figura 4. Comparação do top-k entre as categorias "Frutas"e "Farináceos"usando diferentes funções de similaridade

A Figura 4 mostra os resultados alcançados para duas categorias específicas, especialmente selecionadas por permitirem observar o comportamento discrepante das funções de similaridades escolhidas. Em primeiro lugar, observa-se que no top-1, as duas funções conseguem recuperar strings relevantes. O desempenho cai conforme o valor de k aumenta. Por exemplo, no top-5, menos de 63% das frutas retornadas pela função ED, e menos de 65% dos farináceos retornados pelas função LCS, são relevantes.

Convém destacar que a função ED apresenta resultados melhores para a categoria de "Farináceos", enquanto a função LCS se sai melhor com a categoria de "Frutas". Há

também casos em que nenhuma das funções apresenta um bom resultado, por exemplo, quando um produto é super-especificado em um estabelecimento ("QJ PARM RAL GROS 50G", por exemplo) e sub-especificado em outro ("QUEIJO", por exemplo). Isso demonstra que a forma de descrever produtos dentro do nicho de e-commerce estudado exige o uso de funções de similaridade mais sofisticadas, capazes de detectar nuances onde as funções mais genéricas falham. Por exemplo, as funções usadas não conseguem dar o devido peso à strings curtas, identificando quando elas realmente se referem à abreviações ou prefixos.

5. Considerações Finais

Como visto, a descrição dos produtos em notas fiscais de estabelecimentos dos setor supermercadista é, de forma geral, problemática e insuficiente. Em análises simples já é possível observar diversas divergências de representação, como erros ortográficos, abreviaturas, prefixação, super-especificação e sub-especificação. Tal variação exige o uso de estratégias mais sofisticadas de identificação da similaridade, em relação às funções normalmente usadas para este fim.

Como trabalho futuro, pretende-se investigar o problema mais a fundo identificando/criando funções de similaridade mais abrangentes, e que sejam preferencialmente métricas, de modo a serem aproveitadas dentro de mecanismos de indexação eficientes. A clusterização e rotulagem completamente automatizada também são temas que merecem uma investigação aprofundada, e podem contribuir para a criação de soluções de integração de produtos do setor supermercadista.

Referências

Bergamaschi, S., Guerra, F., and Vincini, M. (2002). A data integration framework for e-commerce product classification. In Horrocks, I. and Hendler, J., editors, *The Semantic Web — ISWC 2002*, pages 379–393, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bergroth, L., Hakonen, H., and Raita, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48.

Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., and Flett, A. (2001). Product data integration in b2b e-commerce. *IEEE Intelligent Systems*, 16(4):54–59.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Neto, H. D. A. and Martinez, A. L. (2016). Nota fiscal de serviços eletrônica: uma análise dos impactos na arrecadação em municípios brasileiros. *Revista de Contabilidade e Organizações*, 10(26):49–62.

Vieira, P. A., Pimenta, D. P., Cruz, A. F. d., and Souza, E. M. S. d. (2019). Efeitos do programa de nota fiscal eletrônica sobre o aumento da arrecadação do estado. *Revista de Administração Pública*, 53:481–491.

Zhang, Y. (2014). Data integration in b2b e-commerce. In *Materials Science, Computer and Information Technology*, volume 989 of *Advanced Materials Research*, pages 4802–4805. Trans Tech Publications Ltd.