ARTIGO

DOI: https://doi.org/10.21728/p2p.2024v11n1e-7124

Data de submissão: 06/07/2024 Data de aprovação: 03/09/2024 Data de publicação: 05/07/2024

PROPOSTA DE METADADOS PARA DESCRIÇÃO DE PRODUTOS DA NOTA FISCAL DE CONSUMIDOR ELETRÔNICA (NFC-*E*) USANDO APRIORI

Diana Maria da Camara Gorayeb[1]
Universidade de Brasília
diana.gorayeb@aluno.unb.br

Claudio Gottschalg-Duque[2]
Universidade de Brasília
klauss@unb.br

Resumo

Este estudo tem como objetivo utilizar o algoritmo Apriori para dar agilidade na busca de termos sem a predefinição pelo especialista, considerando a análise da frequência e da relevância para propor um conjunto com elementos mínimos de metadados que permitam a descrição inequívoca do produto na NFC-*e*. Partiu da problemática de encontrar, por meio de mineração dos dados frequentes, o termo que identifica de forma inequívoca o produto de interesse e quais as relações mais comuns e mais importantes dispostas no campo descrição. Buscaram-se dados da NFC-*e* da Secretaria de Fazenda do Estado do Amazonas, disponibilizados em arquivo .csv, tipo texto, no período de 01/02/2023 a 31/05/2023. A metodologia aplicada se fez a partir da aplicação de sete etapas: seleção do produto de interesse; análise dos dados; seleção da amostra; aplicação do algoritmo Apriori; análise dos termos extraídos do Apriori; obtenção da lista final de termos; análise dos termos finais; proposta de definição dos metadados. A base teórica apoia-se nos estudos da Ciência da Informação, com foco nos campos da mineração de dados, mineração de textos, processamento da linguagem natural, metadados. Como resultado, tem-se, em síntese, a proposta dos metadados em quantidade (quatro) e qualidade: nome do produto, nome da marca comercial do produto, capacidade e tipo da embalagem compõe uma expressão com significado suficiente para o entendimento do conceito que se deseja alcançar no domínio da cerveja.

Palavras-chave: Ciência da Informação; metadados; nota fiscal eletrônica; Apriori-algoritmo.

METADATA PROPOSAL FOR PRODUCT DESCRIPTIONS OF ELECTRONIC CONSUMER INVOICE (NFC-E) USING APRIORI

Abstract

This study aims to use the Apriori algorithm to speed up the search for terms without predefinition by the specialist, considering the analysis of frequency and relevance to propose a set of minimum metadata elements that allow the unambiguous description of the product in the NFC-e . It started with the problem of finding, through frequent data mining, the term that unequivocally identifies the product of interest and which are the most common and most important relationships arranged in the description field. NFC-e data from the Amazonas State Finance

[1] Doutorando no Programa de Pós-Graduação em Ciência da Informação da Universidade de Brasília (PPGCINF/UnB). Possui graduação em Engenharia Elétrica pela Universidade Federal do Amazonas (UFAM) e mestrado em Engenharia Elétrica pela Universidade de São Paulo (USP). Atualmente é professora da Universidade do Estado do Amazonas (UEA) no curso de Engenharia da Computação. Atua como Perito Criminal da Polícia Civil do Estado do Amazonas, área de computação.

[2] Professor Associado da Escola de Ciência da Informação da Universidade de Brasília (FCI-UnB), membro efetivo do Programa de Pós-Graduação em Ciência da Informação (PPGCINF/FCI-UNB) e colaborador da Universidade Federal de Minas Gerais no Programa de Pós-Graduação em Gestão e Organização do Conhecimento (PPGGOC). Licenciado em Letras Modernas, com habilitação em Português e Alemão, pela Faculdade de Letras da Universidade Federal de Minas Gerais (1994), mestre em Psicolinguística pela do Programa de Pós-Graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais (1998), Doutorado-Sanduíche em linguística computacional na Angewandte Sprachwissenschaft und Computerlinguist - Justus-Liebig-Universität Giessen (2003/2004), e Ph.D. em Produção e Gestão da Informação no Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais (2005).

Department was sought, made available in a .csv file, text type, from 02/01/2023 to 05/31/2023. The methodology applied was based on the application of seven steps: selection of the product of interest; data analysis; sample selection; application of the Apriori algorithm; analysis of terms extracted from Apriori; obtaining the final list of terms; analysis of final terms; proposal for defining metadata. The theoretical basis is based on Information Science studies, focusing on the fields of data mining, text mining, natural language processing, metadata. As a result, we have, in summary, the proposal of metadata in quantity (four) and quality: product name, name of the product's commercial brand, capacity and type of packaging compose an expression with sufficient meaning to understand the concept that if you want to achieve in the field of beer.

Keywords: Information Science; metadata; eletronic invoice; Apriori-algorithm.

PROPUESTA DE METADATOS PARA LA DESCRIPCIÓN DE PRODUCTO DE FACTURA ELECTRÓNICA AL CONSUMIDOR (NFC-E) USANDO APRIORI

Resumen

Este estudio pretende utilizar el algoritmo Apriori para agilizar la búsqueda de términos sin predefinición por parte del especialista, considerando el análisis de frecuencia y relevancia para proponer un conjunto de elementos mínimos de metadatos que permitan la descripción inequívoca del producto en el NFC-e. Se partió del problema de encontrar, a través de una frecuente minería de datos, el término que identifique inequívocamente el producto de interés y cuáles son las relaciones más comunes e importantes dispuestas en el campo de descripción. Se buscaron datos NFC-e de la Secretaría de Finanzas del Estado de Amazonas, disponibles en archivo .csv, tipo texto, del 01/02/2023 al 31/05/2023. La metodología aplicada se basó en la aplicación de siete pasos: selección del producto de interés; análisis de los datos; selección de muestras; aplicación del algoritmo Apriori; análisis de términos extraídos de Apriori; obtener la lista final de términos; análisis de términos finales; propuesta para definir metadatos. La base teórica se basa en estudios de Ciencias de la Información, enfocándose en los campos de minería de datos, minería de textos, procesamiento del lenguaje natural, metadatos. Como resultado tenemos, en resumen, la propuesta de metadatos en cantidad (cuatro) y calidad: nombre del producto, nombre de la marca comercial del producto, capacidad y tipo de embalaje componen una expresión con significado suficiente para entender el concepto que si queremos conseguir en el campo de la cerveza.

Palabras clave: Ciencias de la Información; metadatos; factura electrónica; Apriori- algoritmo.

1 INTRODUÇÃO

Este estudo de caso tem como objetivo utilizar o algoritmo Apriori para dar agilidade na busca de termos sem a predefinição pelo especialista, considerando a análise da frequência e da relevância para propor um conjunto com elementos mínimos de metadados que permitam a descrição inequívoca do produto na NFC-*e*.

A problemática reside em encontrar, por meio de mineração dos dados frequentes o termo que identifica de forma inequívoca o produto de interesse e quais as relações mais comuns e mais importantes que ele possui e que estão dispostas no campo descrição, para, assim, caracterizar um conjunto de metadados do produto que associados entre si identificam as transações de NFC-*e* que possuem interesse informacional para a fiscalização do ICMS.

A descrição de produtos no documento da Nota Fiscal Eletrônica (NF-*e*) e na Nota Fiscal de Consumidor Eletrônica (NFC-*e*) é considerada uma tarefa instituída em leis e normas que regulamentam a cobrança e o controle de fiscalização do ICMS em todo território nacional (Frossard, 2011). Contudo, as obrigações regulatórias não descrevem de forma explícita quais elementos são suficientes em qualidade e quantidade para descrever um produto comercializado e permitem que esta descrição seja feita em um campo da nota fiscal, livre de parâmetros do ponto de vista informacional, lógico e computacional.

O conjunto das descrições de produtos está em uma gigantesca base de dados, repleta de informações sobre os produtos e os contribuintes que o comercializam, há grande dificuldade para tratamento e consulta dos dados. Além do volume com o grande número de documentos, outros problemas como inconsistência e a falta de metadados, existência de polissemia e falta de padronização da terminologia para a categorização na descrição dos produtos como uso de abreviações e descrições ambíguas demandam esforços para compreensão e controle da informação.

O estudo utilizou o algoritmo Apriori para dar agilidade na busca de termos sem a predefinição pelo especialista, os dados da NFC-*e* da Secretaria de Fazenda do Estado do Amazonas (SEFAZ/AM) foram disponibilizados em arquivo **.csv**, tipo texto, no período de 01/02/2023 a 31/05/2023. A metodologia aplicada se fez a partir da aplicação de sete etapas, para extração e classificação dos padrões de interesse, na sequência destacada abaixo:

1.ª etapa: Seleção do produto de interesse; Recebimento dos dados.

2.ª etapa: Análise dos dados

3.ª etapa: Seleção da amostra; Extração do *Token* e definição do *itemset.*

4.ª etapa: Aplicação do algoritmo Apriori.

5.ª etapa: Análise dos termos extraídos do Apriori.

6.ª etapa: Obtenção da lista final de termos.

7.ª etapa: Análise dos termos finais; Proposta de definição dos metadados.

A base teórica que orientou este estudo pauta-se nos estudos da Ciência da Informação, com foco nos campos da mineração de dados, mineração de textos, processamento da linguagem natural, metadados. Esse arcabouço teórico traz o conhecimento necessário para que se possa ampliar os conhecimentos nesta área de estudo.

2 CIÊNCIA DA INOFRMAÇÃO: O USO DE METADADOS

O surgimento da Ciência da Informação (CI) está diretamente ligada à ideia de recuperação da informação, por conta da imensa quantidade e variedade de informações que passam a existir a partir da década de quarenta e mais recentemente, por causa dos inúmeros documentos produzidos em meios digitais.

Buckland (1991) apresenta o conceito de informação como coisa, processo e conhecimento, pois pretende alcançar tudo o que é potencialmente informativo, considerando que não somente livros são os responsáveis pela informação, abrindo portas para que qualquer objeto catalogado em instituições possa ser fonte de informação. Por certo que esse entendimento redimensiona as fontes de informação e traz nova perspectiva para estudos na linguagem computacional.

Os objetos não são, por si só, informativos eles dependem das pretensões do conhecimento dos sistemas informacionais. Segundo Capurro e Hjorland (2003) é necessário que outras linguagens, como as palavras sejam utilizadas para ajudar a definir um termo ou representar algo que se pretende alcançar. Assim, para que se tenha o uso real de um termo, é preciso que a definição e as relações entre eles sejam bem definidas. Por isso, na Teoria do Significado de Wittgenstein (1958), ele propõe a definição de termos a partir do uso e do emprego na realidade das pessoas, ou seja, dependentes do contexto, do emprego deles no cotidiano.

Definir um termo é ter domínio da especialidade que ele abrange. Considera-se, de acordo com Lara (2004), que “é um signo linguístico que difere da palavra, unidade da língua geral, por ser qualificado no interior de um discurso de especialidade”. Dessa forma, um termo será considerado de acordo com o uso, o emprego dele no contexto específico e sua especialidade.

Diante do metadados, pode-se relacionar à documentação manual ou eletronicamente de algum objeto informacional extraído por intermédio de linguagens naturais, artificiais, registros textuais, sonoros, imagens, suportes manuscritos, impressos e eletrônicos (Victorino; Pinheiro; Santos, 2015, p. 235). Autores como Rosenfeld, Morville e Arango (2015), Garshol (2004) definem metadados como termos cuja finalidade é descrever e representar objetos como documentos, pessoas, processos, módulo de conteúdo e organizações.

Ampliando o conceito de metadados, Alves (2010) o considera como dados utilizados para representar e identificar um objeto do mundo real, de interesse geral por meio de um recurso informacional. Além disso, Carvalho (2013) evidencia que os metadados podem servir para: criação de catálogos descritivos e operacionais, validar direitos de acesso às informações de autenticação, avaliar conteúdo, melhorar mecanismos de busca e extrair informações. Por conseguinte, os metadados podem servir para uma seleção padrão de estrutura de dados, um conjunto de termos essenciais para a consulta de produtos e serviços.

Por meio dos metadados, a representação do objeto atinge uma correspondência tão perfeita que os termos são classificados como idênticos ou análogos (International..., 2011). As equivalências se dão nas seguintes situações: os termos são sinônimos; os termos são quase-sinônimos; o termo é considerado como desnecessariamente específico e é representado por outro termo com escopo mais amplo; o termo é considerado como desnecessariamente específico e é representado por uma combinação de dois ou mais termo, conhecido como “equivalência composta”.

Assim, a definição de um objeto pode ser feita por uma associação de termos significativos que, uma vez relacionados, irão apresentar conteúdo suficiente para expressar a representação do que se precisa alcançar como um conceito estruturado do ponto de vista informacional e lógico.

A Ciência da Informação, então, no contexto do metadados contribui para ampliar as competências quanto à organização e representação de dados e informações. Nessa perspectiva, são favorecidos os serviços de coleta, registro, filtragem, classificação e entrega de dados e seus metadados. Com isso, efetiva-se o caráter interdisciplinar da Ciência da Informação, como aliada da Ciência da Computação, com possibilidades amplas para tomadas de decisão mais consistentes.

2.1 MINERAÇÃO DE DADOS, MINERAÇÃO DE TEXTOS POR PADRÕES DE INTERESSE OU FREQUENTES

A aquisição do conhecimento para auxílio na tomada de decisão passa pela etapa de busca da informação com entendimento e interpretação dos dados coletados em determinado domínio. A Mineração de Dados (MD) é um dos componentes da Descoberta do Conhecimento em Bases de Dados, conhecida como *Knowledge Discovery in Data Bases* (KDD) e consiste na aplicação de algoritmos de inteligência artificial para exploração de quantidades massivas de dados (Afonso; Duque, 2020, p. 2).

Duas outras atividades são associadas à MD: Pré-processamento e Pós-processamento (Fayyad; Piatetsky-Shapiro; Smyth, 1996) suas técnicas se baseiam em algoritmos como Redes Neurais, usa Modelos Estatísticos e Probabilísticos para o tratamento dos dados (Goldschimidt; Passos, 2005). As tarefas da MD são predição e descrição (Fayyad; Piatetsky-Shapiro; Smyth, 1996) na busca por padrões de interesse e análise de dados, as tarefas se dão na forma representacional de um modelo do tipo classificação, regressão, agrupamento (*clustering*), associação, sumarização, modelagem de sequência, dependência e análise de linhas de tendências.

A MD é obtida pelo conjunto da forma representacional de um modelo, do critério de preferência para sua representação e do método ou algoritmo de busca. A Mineração de Textos (MT) e é considerada como variação da MD, realizada em documentos não estruturados ou semiestruturados com o fito de descobrir padrões e associações relevantes (Goldschimidt; Passos, 2005).

A forma de representação da associação busca identificar as relações que existem ou devem existir entre os dados, seguindo a premissa de “encontrar elementos que implicam na presença de outros em uma mesma transação” (Schuneider, 2002). Em síntese, a associação procura padrões frequentes de associação entre objetos encontrados (Faceli *et al.*, 2021). Destaca-se, também, que os métodos associativos funcionam como tarefas descritivas, interativas, repetitivas e incrementais, cuja realização pode ser feita por meio de algoritmos baseados em Aprendizado de Máquina (AM).

Os métodos associativos utilizam o paradigma dos algoritmos não supervisionados, ou seja, não dependem de um elemento externo para conduzir o aprendizado na extração de um modelo com boa capacidade descritiva. No caso, o aprendizado está voltado para os dados e para o algoritmo de AM, objetivando aprender um modelo, uma regra. Tal regra, no que lhe concerne, deve conseguir alcançar e ser apropriada para novos objetos, desde que estejam no

mesmo domínio em uso e que não façam parte dos dados do treinamento. Esta característica é conhecida como generalização, isto é, a capacidade de eternizar modelo ou regra para novos dados.

Para se realizar a mineração dos dados, Fayyad, Piatetsky-Shapiro e Smyth (1996) indicam seis tarefas:

a) Classificação: momento em que se faz a descoberta de uma função que faça o mapeamento (classificação) de um item de dados em um conjunto de classes pré-definidas;

b) Regressão: quando se faz a descoberta de uma função que mapeie um item de dados em uma variável de predição de valor real;

c) Agrupamento (clusterização): identificação de um conjunto finito de categorias (clusters) que descrevam os dados;

d) Sumarização: momento da busca de uma descrição compacta para um subconjunto de dados;

e) Modelagem de dependência: busca de um modelo que descreva as dependências mais significativas entre as variáveis;

f) Detecção de mudança e desvio: quando se faz a descoberta das mudanças mais significativas nos dados a partir de valores normativos ou previamente medidos.

Neste estudo, o termo Apriori é de grande significância para se alcançar os objetivos propostos. Assim, o termo é definido como um algoritmo de associação, o qual Agrawal e Srikant, em 1994, estabeleceram como algoritmo de descoberta de regras de associação por sua característica de eficiente e responsável mineração de itens frequentes (*itemset*) com finalidade de descobrir o conhecimento em base de dados com várias transações, de modo que cada uma delas suporte um conjunto de itens frequentes (Alpaydin, 2014; Sumithra; Paul, 2010).

Para fins de aplicação do algoritmo considera-se qualquer conjunto de transações assim sintetizadas: T = $\{t_1, t_2, \dots, t_n\}$. Cada transação é combinada por um identificador idt_i e um subconjunto de itens frequentes (*itemset*), identificado como um subconjunto I ⊆ A, sendo A = $\{a_1, a_2, \dots, a_m\}$, universo de m itens. Cada transação "suporta" um subconjunto específico de itens frequentes I. Pode-se afirmar, assim, segundo os autores Katti Faceli *et al*. **(**2021), que dar suporte é "fortalecer ou testemunhar a favor" de um determinado conjunto de itens.

Dessa forma, diz-se que o Suporte de um *itemset* é a fração de transações que o contém, representado por:

$$\sigma_T(I) = \frac{1}{n} |K_T\ (I)|$$

Considerando o conjunto de itens frequentes é viável derivar regras de associação entre eles de natureza probabilística, identificada na forma "*se* antecedente *então então* consequente". Para se chegar ao grau de incerteza da regra verifica-se a confiança da regra, dado em:

$$confiança\ (A\ \ B) = suporte\ (A \cup B)\ /suporte\ (A)$$

De maneira a otimizar a grande quantidade de regras de associação, pode-se extrair um padrão de interesse e de independência da regra, visando selecionar as mais proeminentes. Para isso, baseia-se no princípio de que os padrões que se dão aleatoriamente não são de interesse, o *lift* e convicção seriam assim descritos:

$$lift\ (A\ \ B) = \ confiança(A\ \ B)/suporte(B)$$

$$convicção\ (A\ \ B) = \ [1 - suporte\ \ (B)]/[1 - confiança(A\ \ B)]$$

Dessa forma, o Apriori considera como princípio: se um conjunto de itens é constante, qualquer subconjunto seu também será. Seu inverso é conjuntamente verdadeiro, se um conjunto de itens não é constante, qualquer sucessor seu não o é. Por isso é que o algorítmo utiliza somente os conjuntos de itens constantes de k – 1 elementos para encontrar os conjuntos constantes de k elementos, ampliando sua eficiência.

2.2 PROCESSAMENTO DA LINGUAGEM NATURAL

A literatura indica que o Processamento da Linguagem Natural (PLN) é o conjunto de técnicas computacionais com finalidade de processar texto que aceita uma comunicação entre pessoas e máquinas (Chandra *et al*., 2022). Para geração e recuperação automática de texto, PLN usa o ramo da linguística como: fonologia, morfologia, sintaxe, semântica e pragmática aliadas à inteligência computacional para que "[...] máquinas sejam capazes de ler, escrever e traduzir textos" (Duque, 2005). O autor destaca a contribuição das técnicas de PLN para identificação e extração de sintagmas nominais, sequências de palavras e importantes descritores para recuperar a informação.

Por meio da PLN é possível revelar a significância contextual das palavras usadas nas frases *Relevant Words Recognition (*RWR) (MBoli *et al*., 2021) e assim identificar e classificar entidades nomeadas, além de outros componentes com valor sintático e/ou semântico significativo em textos. O estudo explica que tal feito é viável por meio de técnicas como extração de *tokens*, reconhecimento de entidade nomeada ou *Named Entity Recognition* (NER), desambiguação de entidade nomeada ou *Named Entity Disambiguation* (NED), análise

fragmentada ou superficial conhecida como *Shallow Parsing* e marcação gramatical de partes do discurso conhecida com *Parts-Of-Speech Tagging* (POS).

Seguindo esta linha de pensamento, a extração de *tokens* no texto divide dados textuais em componentes menores e significativos, em unidades de palavras. Para isso, estabelece um processo que usa autômatos finitos e elimina informações e caracteres desnecessários (Cambria; White, 2014). Por meio da técnica NER, é possível extrair menções a entidades nomeadas, já a técnica NED permite a ligação com o significado retirado de uma base de conhecimento e identifica a diferença de significado das palavras com base no contexto.

O *POS Tagging* estipula classificações da estrutura gramatical dos termos a partir da etapa do *DefaultTagger* (Sotaro *et al*., 2016). Além disso, permite compreender não somente os termos individuais, mas as relações entre eles dentro da sentença. Isso porque foi treinado utilizando vasta quantidade de dados linguísticos.

Existem outras perspectivas de aprimoramento da classificação dos termos com o *Rule-Based Tagging* e *Transformation-Based Tagging* (TBT), cujas marcações estão baseadas em regras pré-determinadas, as quais podem, até mesmo, alterar a classificação de um termo sujeitando-se às informações contextuais (Cambria; White, 2014). Já o *Statistical POS Tagging* é uma técnica da linguística computacional, cujas categorias gramaticais são postas em palavras do texto. Essa técnica basea-se em modelos probabilísticos e em aprendizado de máquina com algoritmos *como Conditional Random Fields* (CRF) e *Hidden Markov Models* (HMMs) (Sotaro *et al*., 2016). Outras funções PLN como *Stop Words Removal* possibilitam um filtro para retirar palavras irrelevantes na interpretação textual. Quanto às técnicas do *Stemming* e *Lemmatization*, são filtros para extração dos afixos das palavras reduzindo-as às suas raízes.

2.3 TRABALHOS RELACIONADOS

Estudos de mineração de dados em bases de transações de nota fiscal eletrônica utilizaram técnicas de Inteligência Artificial para classificar o texto de identificar fraudes nos documentos fiscais.

Lenza (2020) tem o objetivo de auxiliar a fiscalização por meio da visualização exploratória dos dados, identificando relações fraudulentas com ocorrências que destoem do padrão. As buscas são feitas sobre as *tags* estruturadas da nota fiscal e utiliza a visão do especialista para a perceção da similaridade ou não dos dados, as técnicas de visualização coordenada incluem gráficos de dispersão e projeções multidimensionais.

Madeira (2015) faz buscas no campo livre da nota fiscal “Discriminação de Serviços” com o objetivo de identificar expressões do tipo “engenharia consultiva”. *Cluster* é a técnica utilizada para classificar o conjunto de registros e a dissonância é avaliada com classificadores baysianos. Outros trabalhos apresentam a aplicação do Apriori com o objetivo de coleta de dados e análise de relações entre atributos em grandes bases de documentos não estruturados.

Chung, Yoo e Choe (2020) aplica o Apriori em bases de dados com prontuários médicos para identificar o risco à saúde e que estão ocultos na anamnese do paciente. As frequências das relações exploradas ajudam na inferência de um contexto oculto relacionado ao risco sem se preocupar com a forma da descrição, ou seja, se o texto corresponde à realidade médica dos pacientes.

Marcelo Maia, Maia e Tsunoda (2020) aplica o Apriori em uma planilha já com os dados transformados contendo solicitações da sociedade registradas pela central com a intensão apoiar a gestão pública identificando relações com um alto valor de confiança que apresente os tipos de reivindicações registradas. A inferência é realizada por meio de busca de palavras-chaves e análise estatística dos resultados.

Em um trabalho similar, a busca de relações com um alto valor de confiança apontada pelo Apriori é relatada no estudo de Costa, Bernardini e Viterbo Filho (2014) sobre base de dados de boletins de ocorrência das rodovias federais e que extraem padrões das causas dos acidentes. O estudo aponta uma comparação entre as regras encontradas pelo Apriori com confiança maior que 0,8 e outros resultados obtidos com algoritmos supervisionados.

Sakai, Nakata e Watada (2018) analisa tabelas de bancos de dados que estão com informações incompletas e busca pelo Apriori inferir atributos para os possíveis valores utilizando conjuntos matemáticos aproximados para recuperação da informação.

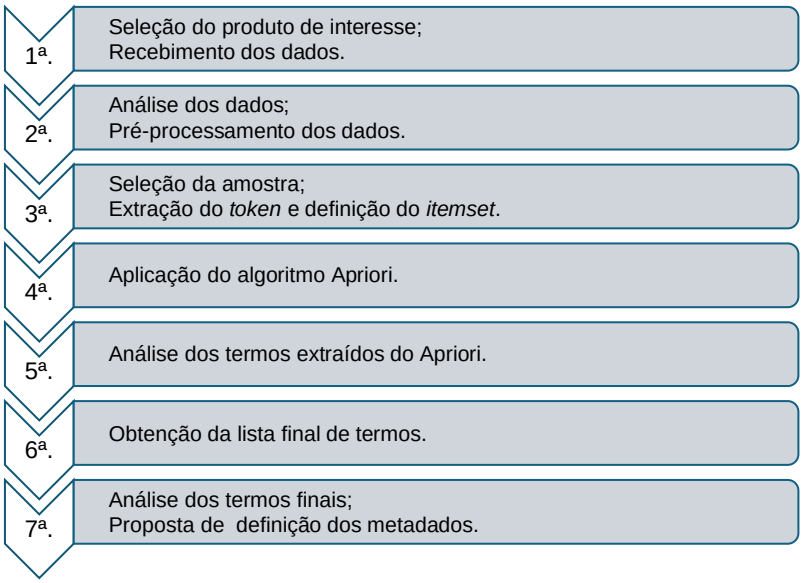
3 PROCEDIMENTOS METODOLÓGICOS

Afonso e Duque (2020, p. 12) após a realização de um estudo com métodos quantitativos evidenciam que os resultados revelam padrões, que por si só, já apresentam conhecimento sobre os dados coletados. Além disso, o uso do método quantitativo pode se dar em menor tempo, no caso dos estudos em mídias sociais e, por realizar filtragem nos dados, gerar apontamentos e indicadores, tornam-se materiais que podem servir para futura análise qualitativa.

Neste estudo, o foco está na busca de dados para extração e classificação dos padrões, sem que uma abordagem qualitativa fosse aplicada. Buscaram-se dados da NFC-*e* da Secretaria de Fazenda do Estado do Amazonas (SEFAZ/AM) disponibilizados em arquivo **.csv**, tipo texto,

no período de 01/02/2023 a 31/05/2023. Foram realizadas sete etapas para extração e classificação dos padrões de interesse, apresentadas na Figura 1.

Figura 1 – Etapas da extração dos termos relevantes e associações para descrição do produto e proposta do modelo de metadados



Fonte: Dados da pesquisa, 2024.

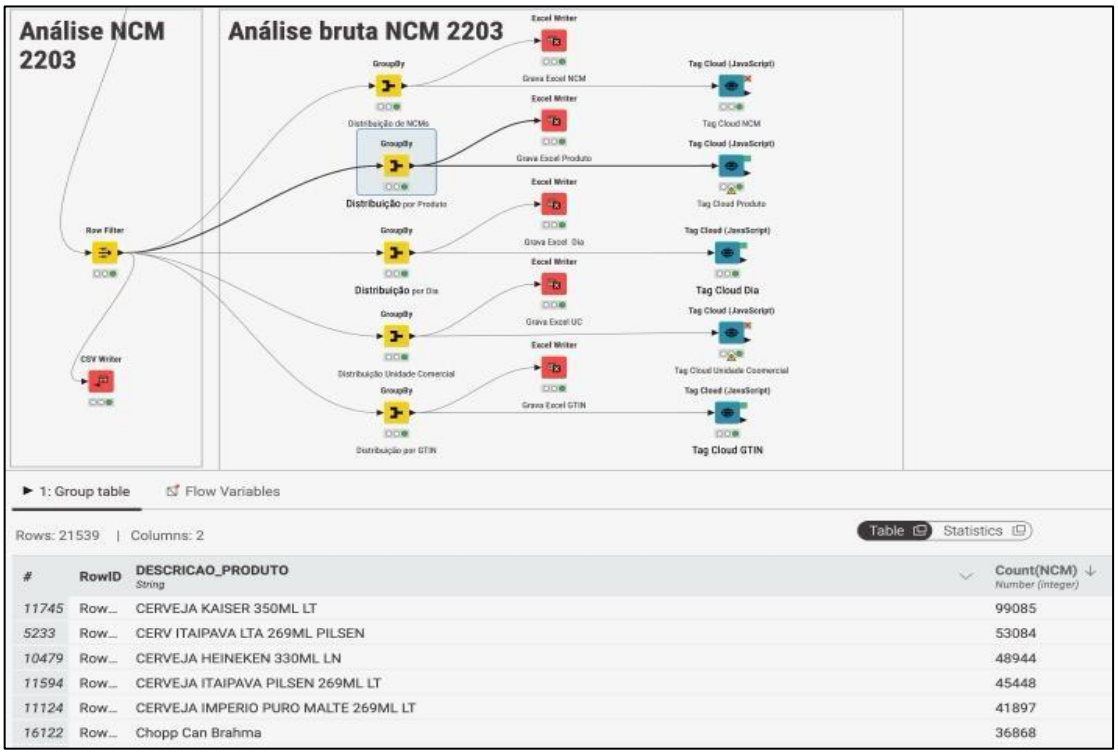
3.1 ETAPAS DA EXTRAÇÃO E CLASSIFICAÇÃO DOS PADRÕES DE INTERESSE DO MODELO DE METADADOS

A primeira etapa tomou como objeto de análise o produto cerveja de malte por meio dos dígitos identificadores do produto: Número Comum do Mercosul (NCM) 2203.xxxx, as subclassificações, os 4 dígitos finais do NCM, não foram relevantes para seleção do dado. A escolha do NCM 2203.xxxx baseou-se na importância estratégica do produto para a Secretaria, além de apresentar inúmeras informações que já se encontravam disponíveis. Isso ajudou na análise da descrição do produto cerveja, considerando legislações e documentos na internet, posto que o estudo foi realizado sem que o especialista da área de fiscalização estivesse presente para predefinir termos ou características do objeto de interesse. Ainda na primeira etapa, os dados foram carregados na plataforma *Knime Analytics* versão 5.2.2, uma ferramenta com código aberto, capaz de realizar a manipulação de dados por meio de algoritmos de AM. A seleção da amostra relevante para NCM 2203.xxxx apresentou 4.019.340 transações.

Na segunda etapa**,** quando se voltou para a análise dos dados, os campos vazios foram eliminados e ajustaram-se os dados via transformações de tipo de dado como: transformar a coluna “NCM” com tipo do dado *number* para tipo *string* para manter a informação tal qual na base original, tipo *char* de 8 posições; transformar na coluna “data de emissão da nota” com o

tipo do dado *string* para tipo *date&time*. Na sequência, os dados foram sumarizados por dia, por NCM com 8 dígitos, por unidade comercial e por descrição do produto para observar o comportamento dos dados selecionados. Na Figura 2, apresentam-se os resultados da descrição dos produtos por volume de transação do NCM 2203.xxxx. Evidenciaram-se, neste caso, os cinco produtos: em 1.º "CERVEJA KAISER 350ML LATA" com 99.085 repetições; 2.º "CERVEJA ITAIPAVA LTA 269ML PILSEN" com 53.084 repetições; 3.º "CERVEJA HEINEKEN 330ML LN" com 48.944 repetições; 4.º "CERVEJA ITAIPAVA PILSEN 269ML LT" com 45.448 repetições e 5.º "CERVEJA IMPERIO PURO MALTE 269ML LT" com 41.897 repetições:

Figura 2 – Resultado preliminar da análise sumarizada dos dados por descrição do produto mais vendidos



#	RowID	DESCRICAO_PRODUTO (String)	Count(NCM) (Number (integer))
11745	Row...	CERVEJA KAISER 350ML LT	99085
5233	Row...	CERV ITAIPAVA LTA 269ML PILSEN	53084
10479	Row...	CERVEJA HEINEKEN 330ML LN	48944
11594	Row...	CERVEJA ITAIPAVA PILSEN 269ML LT	45448
11124	Row...	CERVEJA IMPERIO PURO MALTE 269ML LT	41897
16122	Row...	Chopp Can Brahma	36868

Fonte: Dados da pesquisa, 2024.

Uma NFC-*e* pode apresentar uma ou várias transações no campo descrição do produto correspondendo à venda de um ou de vários produtos como destacado na Figura 1 (venda de 2 produtos de cerveja e correspondente a 2 linhas de transações). Cada linha do resultado acima corresponde a uma transação do campo descrição da NFC-*e* e cada transação se tornará um *itemset* para aplicação do Apriori.

Na terceira etapa**,** foi possível perceber que o resultado de sumarização por descrição dos produtos apresentou produtos que quase nunca se repetem, porque apresentam erro na descrição do produto ou erro na associação do produto ao NCM ou são realmente pouco

comercializados. Nesse caso, não têm valor tributário proeminente para a fiscalização. Quando da análise dos produtos menos comercializados, destacam-se 40 transações do produto "Água sem gás – Água Mineral sem gás 500ml" ao logo do período da amostra como um exemplo do erro na associação do produto com o seu correspondente NCM. Por certo que esses erros geram perdas para o contribuinte, que paga imposto indevidamente sobre a transação, e para a Secretaria, pois acumula informações na base de dados que não são credíveis sobre determinado NCM. Por isso, na etapa de seleção da amostra e definição dos *itemset*, com o intuito de agilizar o processamento, foram retiradas as amostras da "descrição do produto" com número de transações inferiores à 800 repetições. Com essa decisão, reduziu-se a amostra relevante à 2.861.232 transações.

Na quarta etapa o *itemset* serviu de entrada para o algoritmo Apriori. Primeiramente, a intenção é conhecer o Suporte de cada 1 termo do *itemset*. Nessa linha de pensamento, foi estabelecido Suporte mínimo de 0,0000001, um valor baixo visando atingir o maior número de termos. Neste caso, não se estabeleceram as medidas de Confiança e *Lift*. Ao final da execução, o algoritmo apresentou 309 termos distintos que compõem as transações das sentenças de descrição dos produtos. O termo "CERVEJA" com o maior Suporte de valor 0,416, os 70 termos seguintes apresentaram Suporte de valor 0,0000009, como por exemplo os termos: "TAXA", "ENTREGA", "COUVERT", "ARTÍSTICO", "PICOLÉ", "PEPSI", "BANANA", "CARAMELIZADA", "FRITA". Ainda realizando o corte dos termos com menos de 800 repetições, foram encontrados termos não relacionados ao domínio de interesse e estão descritos no produto por erro nas NFC-*e*.

Com vista a confirmar os erros da descrição do produto em relação ao NCM dos NFC-*e*, o Apriori foi executado novamente para apresentar a associação entre 2 termos no *itemset* utilizando Suporte: 0,0001 e Confiança mínima: 0,9. Ao final da execução, encontraram-se 750 regras com os parâmetros acima. Neste conjunto, algumas distorções foram identificadas como: "BANANA", "FRITA" e "CARAMELIZADA". Destacam-se as relações de confiança forte na combinação de 2 termos como: consequente: "BANANA" e antecedente "FRITA" ou "CARAMELIZADA", entretanto essas regras, embora sempre estejam juntas e sejam consideradas regras fortes (*Lift* muito alto: 2.192.609 e Confiança = 1) surgiram com o valor de Suporte praticamente zero, evidenciando que essas relações raramente aparecem, considerando o volume de dados.

Na quinta etapa a análise do Apriori, o processo passa a ser iterativo e incremental, posto que são selecionados apenas os dados que realmente possuem valor para o NCM 2203.xxxx e que podem ser utilizados para representar o produto. É válido destacar que o Apriori foi

executado com as mesmas medidas para apresentar associação entre 3, 4 e até 5 termos do *itemset*, confirmando a existência de termos fora do domínio, por exemplo: para associação de 3 termos existem 5.360 regras e apenas 3 são com os termos “BANANA”, “FRITA” e “CARAMELIZADA”. Nesse contexto, os termos com baixa ocorrência (até 3 vezes) foram retirados da lista uma vez que não faziam parte do domínio de interesse, chegando a um total de mais 39 termos extraídos:

Figura 3 – Termos com baixa ocorrência extraídos da amostra

ITEM	PALAVRA	ITEM	PALAVRA
1	ACCESS	21	FRANGO
2	AGUA	22	FRITA
3	ÁGUA	23	GAS
4	ARTISTICO	24	GORJETA
5	BA	25	GUSTA
6	BANANA	26	NA
7	BORDA	27	PEPSI
8	C/LIMAO	28	PICOLE
9	CALDERETA	29	REFRI
10	CANECA	30	SAL
11	CARAMELIZADA	31	SEM
12	CLARO	32	SUCO
13	COCA	33	SUJO
14	COLA	34	TEMPERADO
15	CONCEDIDA	35	TUL
16	CONVITE	36	VALET
17	COPO	37	XXXXXXXX
18	COUVERT	38	REFRIGERANTE
19	DIVERSAS	39	GÁS
20	DRINK		

Fonte: Dados da pesquisa, 2024.

Ampliando a análise, chega-se ao resultado do Apriori de modo mais consistente, posto que o termo consequente “CERVEJA” demonstra inclinação a se confirmar como termo principal da amostra, visto ser o que mais se repete, afinando-se com a descrição do NCM 2203.xxxx. É prudente destacar que tal coincidência, embora apresente aspecto importante, não é um resultado significativo para o objetivo deste estudo. Por exemplo, em um estudo sobre produtos farmacêuticos com NCM 3003.xxxx e 3004.xxxx, um resultado relevante do Apriori poderia exibir “Dipirona” ou “Paracetamol” como os produtos que mais se repetem, mesmo sem haver coincidência com a descrição dos NCM que é ‘Medicamentos’.

Com a finalidade de verificar a incidência do termo “CERVEJA” para as regras de associação no Apriori para 3, 4 e 5 termos, foi feito um levantamento apresentado na Tabela 1.

Tabela 1 – Resultado das associações para 3, 4 e 5 termos com a descrição CERVEJA como consequente

Termo	Nº de associações	Nº de regras	% de incidência do termo
CERVEJA	3	5.360	14,16% - 759 regras
CERVEJA	4	14.084	12,68% - 1.768 regras
CERVEJA	5	19.745	10,48% - 2.070 regras

Fonte: Dados da pesquisa, 2024.

Como evidenciado, o termo principal CERVEJA varia em um percentual entre 10% e 14% aproximadamente enquanto consequente das regras para associações de 3, 4 e 5 termos. Quanto aos antecedentes dessas mesmas regras, apresentam termos do domínio de cerveja como: “LATA”, “KAISER”, “ITAIPAVA”, “350ML”, “269ML” (Figura 4).

Figura 4 – Associações de 4 termos com o elemento “CERVEJA” no consequente

15658	rule1...	0.004	0.505	1.214	CERVEJA	<---	[LATA,350ML,KAISER]
15656	rule1...	0.004	0.734	1.766	CERVEJA	<---	[SKOL,-,269ML]
15645	rule1...	0.004	1	2.406	CERVEJA	<---	[DEVASSA,PURO]
15635	rule1...	0.004	1	2.406	CERVEJA	<---	[HEINEKEN,-,LONG]
15638	rule1...	0.004	1	2.406	CERVEJA	<---	[HEINEKEN,NECK,-]
15622	rule1...	0.004	0.409	0.983	CERVEJA	<---	[SPATEN,355ML]
15615	rule1...	0.004	0.621	1.494	CERVEJA	<---	[MALTE,BOHEMIA,269ML]
15618	rule1...	0.004	0.779	1.873	CERVEJA	<---	[PURO,BOHEMIA,269ML]
15614	rule1...	0.004	0.448	1.078	CERVEJA	<---	[ANTARCTICA,269ML]
15609	rule1...	0.004	0.429	1.032	CERVEJA	<---	[ANTARCTICA,ORIGINAL]
15606	rule1...	0.004	0.428	1.029	CERVEJA	<---	[SLEEK,350ML,BOHEMIA]
15603	rule1...	0.004	0.476	1.144	CERVEJA	<---	[BUDWEISER,269ML]
15594	rule1...	0.004	0.613	1.475	CERVEJA	<---	[PILSEN,UN,269ML]
15590	rule1...	0.004	0.702	75.972	[CERVEJA]	<---	[12X269ML,ITAIPAVA]

Fonte: Dados da pesquisa, 2024.

É importante notar que os mesmos termos exemplificados acima “LATA”, “KAISER”, “ITAIPAVA”, “350ML”, “269ML”, quando explorados individualmente no papel de consequentes, na grande maioria das vezes, exibem sentenças que descrevem o produto, mas não contêm o termo “CERVEJA”. Com isso, o percentual de regras com o termo “CERVEJA” está restrito aos valores proporcionados pela amostra. A figura abaixo apresenta um exemplo real de descrição do produto sem termo cerveja em NFC-*e*:

Figura 5 – Exemplo de descrição do produto na NFC-*e* sem o termo cerveja

Fonte: Acervo pessoal do autor, 2024.

Na Figura, linha 3, item 003, a associação entre os termos revela uma relação do tipo consequente “HEINEKEN” e antecedentes “LN, 330ML” ou consequente “330ML” e antecedentes “HEINEKEN, LN”, isto é, não apresenta uma expressão compreendendo ou iniciada pelo termo “CERVEJA”. De igual forma, na base estudada outras associações podem ser extraídas para o consequente “330ML”, cujos antecedentes aparecem como “CORONA, LONG” (linha 5046 da Figura 7) ou “STELLA, UN” (linha 5701 da Figura 7), associando entre si termos salientes do domínio da cerveja, sem necessariamente apresentar “CERVEJA” ou “CERV” na descrição da sentença.

Figura 6 – Associações de 3 temos sem o elemento “CERVEJA” ou “CERV”

5046	rule5...	0.001	0.808	5.803	330ML	<---	[CORONA,LONG]
5047	rule5...	0.001	1	7.183	330ML	<---	[CORONA,NECK]
5048	rule5...	0.001	0.808	5.803	330ML	<---	[EXTRA,LONG]
5049	rule5...	0.001	1	7.183	330ML	<---	[EXTRA,NECK]
5216	rule5...	0.002	1	7.183	330ML	<---	[CORONA,UN]
5253	rule5...	0.002	1	7.183	330ML	<---	[EXTRA,-]
5296	rule5...	0.002	0.764	5.491	330ML	<---	[LAGER,STELLA]
5297	rule5...	0.002	1	7.183	330ML	<---	[LAGER,ARTOIS]
5398	rule5...	0.002	1	7.183	330ML	<---	[BUDWEISER,1X330ML]
5701	rule5...	0.002	0.783	5.622	330ML	<---	[UN,STELLA]
5702	rule5...	0.002	0.783	5.622	330ML	<---	[ARTOIS,UN]
5774	rule5...	0.002	1	7.183	330ML	<---	[CORONA,-]
5911	rule5...	0.002	1	7.183	330ML	<---	[HEINEKEN,BEER]

Fonte: Dados da pesquisa (2024).

Ainda na quinta etapa, foi realizada a análise sintática e semântica dos termos frequentes, identificaram-se os ruídos nos dados da amostra, que não contribuiriam na descrição do produto nem na transação: “CERVEJA IMPERIO P MALTE PILSEN269ML-1X269ML”. Nessa sentença, é perceptível que vários ruídos que dificultam a comunicação e interpretação da descrição do produto:

Quadro 1 – Análise de ruídos na transação “CERVEJA IMPERIO P MALTE PILSEN269ML-1X269ML”

1º.	Letra P entre os termos “IMPERIO” e “MALTE”: “IMPERIO P MALTE”
2º.	Duplicidade do termo “269ML” na mesma transação: “1X269ML” e “PILSEN269ML”
3º.	Expressão “PILSEN269ML” formada a partir de 2 termos frequentes com significado para as características do produto e que são “PILSEN” e “269ML”

Fonte: Dados da pesquisa, 2024.

Os ruídos comprometem a lista final dos termos frequentes e relevantes e dificultam a sua classificação. Visando corrigir estes ruídos, também melhorar a qualidade dos termos, aplicaram-se algoritmos de PLN. O primeiro passo para trabalhar com PLN é transformar o texto *String* para

um formato *TextDocument*, de forma a se manter somente as colunas de NCM e a descrição do produto. Na sequência, outros tratamentos do texto foram executados na sequência:

Quadro 2 – Sequência da aplicação de funções PLN na base de dados NFC-*e*

Punctuation Erasure	Para retirar todos os caracteres de pontuação.
N Chars Filter	Para retirar os termos com apenas 1 caractere.
Number Filter	Para retirar alguns caracteres numéricos.
POS	Para fazer a análise sintática do texto.
Tag Filter	Para selecionar as palavras já classificadas pelo *POS* com exceção de determinantes, pronomes possesivos e advérbios: WDT (), WP (), WP$ (), WRB ().
Stop Word Filter	Para retirar palavras que não são necessárias para o significado da descrição do produto.
Snowball Stemmer	Para extrair os radicais das palavras e minimizar o erro de representação da linguagem dos termos.
Bag of Words	Para criar o *itemset* correspondente a 11.891.415 linhas onde cada linha representa um termo já classificado: o que era 1 linha de sentença de descrição do produto “CERVEJA SKOL LATA 350ML” foi transformada em 4 linhas: 1ª. linha “CERVEJA”, 2ª. Linha “SKOL”, 3ª. Linha “LATA” e 4ª. Linha “350ML”.
Group by e *Row Filter*	Para identificar e eliminar termos que aparecem em menos de 1000 repetições no universo de 11.891.415 linhas.

Fonte: Dados da pesquisa (2024).

Na sexta etapa, restou uma consulta de 11.874.260 radicais dos termos reagrupados na forma de transação da NFC-*e*, totalizando 2.605.761 linhas de descrições de produto, em seguida aplicou o processo de *token* (*cell splitter*) para o Apriori com Suporte mínimo de 0,000001, resultando na lista de 196 radicais dos termos, apresentados na Figura 7.

Figura 7 – Lista com resultado dos 196 termos finais do evento 1 para construção da ontologia

ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO	ITEM	SUPORTE	TERMO
1	0,000132783	cervstell	51	0,000605197	35l	101	0,001438351	refriger	151	0,010912743	la
2	0,000176532	rio	52	0,000617862	ext	102	0,001506278	250ml	152	0,011660701	500ml
3	0,000176532	negr	53	0,000618629	pma	103	0,001511267	bra	153	0,012059049	antart
4	0,000193418	cervaj	54	0,000627072	gfa	104	0,001582647	camar	154	0,012085145	1x350ml
5	0,000238318	mat	55	0,000627072	vd	105	0,001604138	way	155	0,013518124	1x269ml
6	0,000279765	export	56	0,000627072	c23	106	0,001633304	1x355ml	156	0,013597179	355ml
7	0,000293196	cart	57	0,000647795	one	107	0,001634455	cerp	157	0,015648787	teor
8	0,000308931	12x1l	58	0,000662762	cervgarraf	108	0,001652492	c15	158	0,016020656	cor
9	0,000308931	ambev	59	0,000663914	ex	109	0,001720035	sens	159	0,018260692	arto
10	0,000309315	eisenbah	60	0,000663914	transp	110	0,001802161	dm	160	0,020283518	stell
11	0,000309315	pil	61	0,000674275	itaipav	111	0,001824803	kg	161	0,021969398	crystal
12	0,000313536	unic	62	0,000686172	hour	112	0,001918825	litra	162	0,021978608	600ml
13	0,000321595	269mlx15un	63	0,000710349	cervpur	113	0,001954899	npal	163	0,026428748	300ml
14	0,00032927	malzbi	64	0,000719176	343ml	114	0,002037792	glacial	164	0,027919291	original
15	0,000335027	antarctic	65	0,000739899	und	115	0,002060051	269m	165	0,029793216	sleek
16	0,000339632	12un	66	0,000787486	350g	116	0,002149852	beer	166	0,030731521	neck
17	0,00034654	caracu	67	0,00079708	gold	117	0,002230442	cx-12	167	0,033220622	lag
18	0,000354215	golden	68	0,00079708	210ml	118	0,002257306	beats	168	0,036062402	antarct
19	0,000354215	ale	69	0,000813198	boh	119	0,00228724	sle	169	0,036230875	long
20	0,000354599	wit	70	0,000846969	cristal	120	0,002352096	brasil	170	0,036479171	bohem
21	0,000365344	cx12und	71	0,000856947	pr	121	0,002375122	imper	171	0,038325464	spaten
22	0,000373403	cervbohem	72	0,000860401	cer	122	0,002510207	baden	172	0,040162548	amstel
23	0,000386451	unfiltered	73	0,00088573	12x350	123	0,002635698	gf	173	0,042637448	
24	0,000386451	teo	74	0,00089187	ita	124	0,002748909	c12	174	0,045929385	imperi
25	0,000391057	cervbrahm	75	0,00089187	pc12	125	0,002943478	cx	175	0,05199671	dupl
26	0,000391824	dev	76	0,000909523	gás	126	0,003523347	pm	176	0,054653516	budweis
27	0,000392208	eisen	77	0,000915663	spat	127	0,004002286	sh	177	0,064056911	ml
28	0,000392208	unfil	78	0,000917582	1l	128	0,004111659	happy	178	0,068274872	lat
29	0,000397964	lo	79	0,000923722	fi	129	0,004363409	premium	179	0,070218642	skol
30	0,000399883	antar	80	0,00092449	pmalt	130	0,004651232	subzer	180	0,073189751	pur
31	0,000418304	cx15	81	0,00093677	heinek	131	0,004701122	petr	181	0,075617833	un
32	0,000440179	cx-24	82	0,000951737	pils	132	0,004971292	lt350ml	182	0,080941805	lta
33	0,000449389	1x600ml	83	0,000990114	bud	133	0,005132474	12x350ml	183	0,08136318	chopp
34	0,000453994	350mlgross	84	0,001025036	munichpur	134	0,005304784	proib	184	0,099792728	kais
35	0,000454762	lokal	85	0,001036166	longneck	135	0,005384224	cervitaip	185	0,110621043	ln
36	0,000461669	stel	86	0,001061494	cervskol	136	0,005396888	chop	186	0,124039388	itaip
37	0,000461669	art	87	0,001097568	trig	137	0,005691619	eisenbahn	187	0,13225503	brahm
38	0,000502348	lata	88	0,001124432	ma	138	0,006255754	dobr	188	0,137264315	heineken
39	0,00052192	300mlchopp	89	0,00117317	dmalt	139	0,006263813	orig	189	0,138077514	330ml
40	0,000523072	269mll	90	0,001175089	prem	140	0,006268035	devass	190	0,163089017	pilsen
41	0,000536504	dup	91	0,001183915	6x330ml	141	0,006286455	tijuc	191	0,163785934	malt
42	0,000537271	gt	92	0,001214616	cozumel	142	0,006509039	ow	192	0,233843395	269ml
43	0,000538422	medi	93	0,001225362	350m	143	0,006806457	schin	193	0,282860938	lt
44	0,000539574	cho	94	0,001244934	congel	144	0,00692504	12x269ml	194	0,297839288	350ml
45	0,000558378	witbi	95	0,001244934	maring	145	0,007370592	tig	195	0,397935958	cerv
46	0,000567205	ar	96	0,001281392	lon	146	0,007453101	lt269ml	196	0,4305084	cervej
47	0,000570659	grf	97	0,001323606	330355l	147	0,008120085	can			
48	0,000579869	12x269	98	0,001350853	lneck	148	0,008128911	1x330ml			
49	0,000583707	ipa	99	0,001386543	becks	149	0,009698127	munich			
50	0,000585625	origin	100	0,001389229	litr	150	0,010754632	extra			

Fonte: Dados da pesquisa, 2024.

Na sétima etapa**,** os termos foram organizados a partir da equivalência (Quadro 3).

Quadro 3 – Organização e agrupamento dos radicais dos termos a partir do significado

Organizado por	Termos	Quant
Nome do produto	cervej, cerv, cervstell, cervaj, cervbohem, cervbrahm, cerevgarraf, cervpur, cervskol.	9
Quantidade do produto embalado	350ml, 269ml, 330ml, ml, 300ml, 600ml, 355ml, 500ml, 350mlgross, 300mlchopp, 269mll, 300355l, 350m, 6x330ml, 210ml, 350g, 343ml, 35l, 250ml, 269m, 1l, litr, lt269ml, lt250ml, lt269m.	25
Quantidade do produto vendido	1x269ml, 1x350ml, 1x600ml, 12x1l, 269mlx15un, 12un, cx12und, cx15, cx-24, 12x269, pc12, 12x350, cx23, 1x355ml, c15, cx-12, c12, 12x269, 12x269ml, 1x330ml, 12x350ml, lt350ml, c23, 12x269ml.	24
Unidade relacionada ao produto vendido	un, cx, gfa, gf, und.	5
Nome da marca do produto comercializado	heineken, brahm, itaip, kais, ita, skol, budweis, imperi, amistel, spaten, bohem, antarct, original, stell, arto, antart, ambev, eisenbah, malzbi, antarctic, caracu, eisen, antar, stel, orign, itaipav, cozumel, becks, munich, tijuc, devass, schin, orig, eisenbahn, proib, petr, baden, brasil, imper, beats, cerp.	41
Forma do recipiente do produto	lt, ln, lat, long, neck, la, lata, longneck, lon, lneck.	10
Características complementares	malt, pilsen, chopp, pur, dupl, lag, crystal, extra, mat, export, cart, pil, golden, dup, transp, gold, trig, prem, teor, beer.	20
Total		134

Fonte: Dados da pesquisa, 2024.

Assim, dos 196 termos iniciais, 134 termos, ou seja, 68,36% estão organizados pela equivalência de representação e significado, 62 termos (31,64%) não são pertinentes ao domínio cerveja como: “cor”, “refrig”, “way”, “dm”, “kg”, “pm”, “sh”, “ow”, “can”, “pma”, “rio”, “negr”, “gt”, “grt”, “ar” etc. A partir da organização dos termos é possível avaliar:

Quadro 4 – Análise sobre o agrupamento gerado

a)	A quantidade de produto vendido e a unidade, embora estejam, algumas vezes, expressos no texto do campo descrição, não fazem parte das características do produto. Com vista a isso a legislação da nota fiscal estabelece que essas informações sejam descritas em outros dois campos da NFC-*e* “quantidade” e “unidade” ao lado de outros campos estruturados do ponto de vista computacional na base de dados NFC-*e*: código, valor unitário e valor total;
b)	O nome do produto é a representação do que se quer alcançar e, portanto, ponto de partida para uma descrição pois existe em e por si mesmo;
c)	A quantidade de produto embalado está descrita frequentemente na descrição do produto, pois não há outro campo para registro e representa a caraterística do volume do líquido embalado em um recipiente ou a sua capacidade;
d)	Nome da marca do produto comercializado complementa o termo do nome do produto “cerveja”, caracterizando o nome comercial vinculado ao produto para a venda;
e)	Forma do recipiente do produto representa a forma e material de produção, descreve o tipo de embalagem para o transporte do produto vendido;
f)	Características complementares são termos que complementam o nome da marca do produto, evidenciando características de produção do produto.

Fonte: Dados da pesquisa, 2024.

Após ter os resultados, tornou-se notável que a descrição do produto cerveja detém 4 características qualitativas relevantes e frequentes: nome do produto, nome da marca comercial do produto, capacidade e tipo da embalagem. Desse modo, entende-se que tais características podem representar os metadados da descrição do produto, porém a expressão que melhor comunica o produto considerando o domínio de interesse é: “Nome_Produto + Marca_Produto + Capacidade_Embalagem + Tipo_Embalagem”. A sentença pode ser descrita na seguinte forma: [“CERVEJA”, ou “CERV”] + [“ORIGINAL” ou “HEINEKEN” ou “SKOL”] + [“269ML” ou “330ML” ou “350ML”] + [“LN” ou “LATA”].

As características complementares são informações familiares no domínio da cerveja e podem acrescentar determinadas qualidades para descrever tipos de produção do item cerveja, por exemplo: puro malte, *golden* ou *export*. Entretanto, a comunicação e percepção da descrição do produto não depende dessas expressões, logo não é critério para o significado e uso correto do produto.

4 APRESENTAÇÃO DOS RESULTADOS

Após seguir as sete etapas, chegou-se a um resultado preliminar, o qual aponta para alguns resultados preliminares, descritos a seguir:

1. Os termos relevantes para o domínio e encontrados na lista final foram selecionados pela análise da frequência e pela forma de associação com que aparecem nas transações. Contudo, outros filtros poderiam ser aplicados para o refinamento da informação como: extrair termos com até 2 caracteres, sem o prejuízo da descrição do produto: “dm”, “kg”, “pr”, “sh”, “pm”, “fi”;

2. O significado dos termos considerando seu uso real no domínio de interesse da cerveja pode ser confirmado pelos documentos regulatórios da Secretaria ou dicionários de domínio público: “LTA” ou “LT” ou “LATA” – significado: “folha fina de ferro estanhado”; “folha de flandres”; “recipiente de folha de flandres para uso doméstico e industrial, principalmente para acondicionamento de conservas e líquidos, tais como óleo, gasolina, tintas, água etc.” (Lata, 2015).

3. A organização dos termos foi feita a partir da análise sintática e equivalência na forma de representação e significado: são sinônimos, quase-sinônimos; são representados por outro termo ou são representados por um conjunto de termos; tem a equivalência composta. Um exemplo de termos análogos, mas com erro na expressão da linguagem: “CERVEJA”, “CERV”, “CERVEJ” etc.;

4. Uma vez organizados os conjuntos de termos, foram classificados pelo uso real no domínio cerveja como características que existem por si e em si e quando associadas umas às outras propõem um padrão para seu uso correto, um critério para o entendimento da descrição do produto. A classificação propõe uma estrutura de tipo de dados com poder de promover a descrição do produto a partir do nome “CERVEJA” ou equivalente, caracterizando-o com as qualidades que mais o descrevem e que se assemelham ao tratamento dispensado pela Secretaria, formando um vocabulário especializado;

5. A proposta dos metadados em quantidade (quatro) e qualidade: nome do produto, nome da marca comercial do produto, capacidade e tipo da embalagem compõe uma expressão com significado suficiente para o entendimento do conceito que se deseja alcançar no domínio da cerveja.

5 CONSIDERAÇÕES FINAIS

Com o objetivo de utilizar o algoritmo Apriori para dar agilidade na busca de termos sem a predefinição pelo especialista, considerando a análise da frequência e da relevância, buscou-se propor um conjunto com elementos mínimos de metadados que permitam a descrição inequívoca do produto na NFC-*e*.

Por meio da mineração dos dados frequentes foi proposto o termo que identifica de forma inequívoca o produto de interesse e as relações mais comuns e mais importantes que ele possui e que estão dispostas no campo descrição. A partir disso, caracterizar um conjunto de metadados do produto que, associados entre si, identificam as transações de NFC-e que possuem interesse informacional para a fiscalização do ICMS.

Com a aplicação de metodologia consistente, estruturada em campos da Ciência da Informação, foi possível obter resultados satisfatórios, quais sejam:

Os termos relevantes para o domínio, presentes na lista final foram selecionados pela análise da frequência e pela forma de associação, mas outros filtros também poderiam ser aplicados para o refinamento da informação, como termos com até 2 caracteres, sem o prejuízo da descrição do produto. O significado dos termos tendo em vista seu uso real no domínio de interesse da cerveja pode ser confirmado pelos documentos regulatórios da Secretaria ou pelos dicionários de domínio público: “LTA” ou “LT” ou “LATA”. A organização dos termos foi feita com base na análise sintática e na equivalência na forma de representação e significado, averiguando se são sinônimos, quase-sinônimos; se são representados por outro termo ou se são representados por um conjunto de termos; se tem a equivalência composta, como ocorre em termos análogos, mas com erro na expressão da linguagem: “CERVEJA”, “CERV”, “CERVEJ” etc. A classificação propõe uma estrutura de tipo de dados com poder de promover a descrição do produto a partir do nome “CERVEJA” ou equivalente, caracterizando-o com as qualidades que mais o descrevem e que se assemelham ao tratamento dispensado pela Secretaria, formando um vocabulário especializado.

Em suma, a proposta dos metadados em quantidade (quatro) e qualidade: nome do produto, nome da marca comercial do produto, capacidade e tipo da embalagem compõe uma expressão com significado suficiente para o entendimento do conceito que se deseja alcançar no domínio da cerveja.

REFERÊNCIAS

AFONSO, Alexandre Ribeiro; DUQUE, Cláudio Gottschalg. Mineração de textos aplicada a postagens do Twitter sobre Coronavírus: uma análise na linha do tempo. **Liinc em Revista**, Rio de Janeiro, v. 16, n. 2, p. 1-13, dez. 2020. Disponível em: https://revista.ibict.br/liinc/article/view/5325. Acesso em: 28 ago. 2024.

ALPAYDIN, Ethem. **Introduction to Machine Learning.** 3 ed. Massachusetts: MIT Press, 2014.

ALVES, Rachel Cristina Vesú. **Metadados como elementos do processo de catalogação**. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Marília, São Paulo, Universidade Estadual Paulista Júlio Mesquita Filho, 2010. Disponível em: https://repositorio.unesp.br/bitstreams/e1fd7853-781a-46bd-beee-037375fb178f/download. Acesso em: 28 ago. 2024.

BUCKLAND, Michael. Information as Thing. **Journal of the American Society for Information Science**, Nova Jersey, v. 42, n. 5, p. 351-360, jun. 1991. Disponível em: https://ppggoc.eci.ufmg.br/downloads/bibliografia/Buckland1991.pdf Acesso em: 13 jan. 2024.

CAMBRIA, Erik; WHITE, Bebo. Jumping NLP Curves: A Review of Natural Language Processing Research. **IEEE Computational Intelligence Magazine**, [*S. l.*], v. 9, n. 2, p. 48-57, abr. 2014. Disponível em: https://ieeexplore.ieee.org/abstract/document/6786458. Acesso em: 28 ago. 2024.

CAPURRO, Rafael; HJORLAND, Birger. O conceito de informação. **Perspectivas em Ciência da Informação**, Belo horizonte, v. 12, n. 1, p. 148-207, jan./abr. 2007. Disponível em: https://doi.org/10.1590/S1413-99362007000100012. Acesso em: 28 ago. 2024.

CARVALHO, E. O. **Uma proposta de interdisciplinaridade entre arquitetura da informação e ciência da computação:** linguagem SOWL para as ontologias da Web utilizando o formalismo dos grafos conceituais. 2013. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2013.

CHANDRA, Ritesh *et al*. Natural Language processing and Ontology based Decision Support System for Diabetic Patients. *In:* INTERNATIONAL CONFERENCE ON ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION (EECSI), 9., 2022, Indonesia. **Anais** […], Indonesia.: IEEE, 2022. Disponível em: https://ieeexplore.ieee.org/document/9946601. Acesso em: 13 jan. 2024.

CHUNG, Kyungyong; YOO, Hyun; CHOE**,** Do-Eun. Ambient context-based modeling for health risk assessment using deep neural network. **Journal of Ambient Intelligence and Humanized Computing**, Germany, v. 11, p. 1387-1395, set. 2020. Disponível em: https://link.springer.com/article/10.1007/s12652-018-1033-7. Acesso em: 28 ago. 2024.

COSTA, Jefferson de Jesus; BERNARDINI, Flávia Cristina; VITERBO FILHO, José. A mineração de dados e a qualidade de conhecimento extraídos dos boletins de ocorrência das rodovias federais brasileiras. **AtoZ**, Paraná, v. 3, n. 2, p.139-157, dez. 2014. Disponível em: https://revistas.ufpr.br/atoz/article/view/41346. Acesso em: 28 ago. 2024.

DUQUE, Claudio Gottschalg. **SiRILiCO uma proposta para um sistema de Recuperação da Informação baseado em Teorias da Lingüística Computacional e Ontologia**. 2005. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005. Disponível em: http://hdl.handle.net/1843/EARM-7HBND8. Acesso em: 28 ago. 2024.

FACELI, Katti *et al*. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro, Editora LTC, 2021.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory. SMITH, Padhraic. From datamining to knowiedge Discovery. **IA Magazine**, [*S. l.*], v. 17, n. 3, p. 37-54, 1996. Disponível em: https://doi.org/10.1609/aimag.v17i3.1230. Acesso em: 13 jan. 2023.

FROSSARD, Dermeval. **ICMS Genérico.** Rio de Janeiro, Editora Ferreira, 2011.

GARSHOL, Lars Marius. Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all. **Journal of Information Science**, [*S.l.*], v. 30, n. 4, p. 378-391, ago. 2004. Disponível em: https://doi.org/10.1177/0165551504045856. Acesso em: 28 ago. 2024.

GOLDSCHIMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 25964-1**: Information and documentation — Thesauri and interoperability with other vocabularies: part 1: Thesauri for information retrieval. 1. ed. Geneva: ISO, 2011.

LARA, Marilda Lopes Ginez de. Diferenças conceituais sobre termos e definições e implicações na organização da linguagem documentária. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 91-96, dez. 2004. Disponível em: https://doi.org/10.18225/ci.inf.v33i2.1050. Acesso em: 28 ago. 2024.

LATA. *In*: Dicionário Brasileiro da Língua Portuguesa. São Paulo: Editora Melhoramentos, 2015.

LENZA, Frederico de Paiva. **Mineração visual de dados em Notas Fiscais do Consumidor Eletrônicas**. 2020. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2020. Disponível em: https://bdm.unb.br/bitstream/10483/27582/1/2020_FredericoDePaivaLenza_tcc.pdf. Acesso em: 28 ago. 2024.

MADEIRA, Renato de Oliveira Caldas. **Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais**. 2015. Dissertação (Mestrado Matemática Aplicada) – Fundação Getúlio Vargas, Rio de Janeiro, 2015. Disponível em: https://hdl.handle.net/10438/14593. Acesso em: 28 ago. 2024.

MAIA, Marcelo; MAIA, Marcos; TSUNODA, Denise Fukumi. Mineração de dados no apoio a gestão púbica municipal: conhecendo demandas da cidade de Curitiba pela “Central 156”. **Revista Tecnologia e Sociedade**, Curitiba, v. 16, n. 40, p. 91-96, abr./jun. 2020. Disponível em: https://periodicos.utfpr.edu.br/rts/article/view/10335. Acesso em: 28 ago. 2024.

MBOLI, Julius Sechang *et al*. Domain Experts and Natural Language Processing in the Evaluation of Circular Economy Business Model Ontology. *In*: INTERNATIONAL

CONFERENCE ON SEMANTIC COMPUTING (ICSC), 15., 2021. **Anais** [...] California: IEEE, 2021. Disponível em: https://ieeexplore.ieee.org/document/9364548. Acesso em: 28 ago. 2024.

ROSENFELD, Louis; MORVILLE, Peter; ARANGO, Jorge. **Information Architecture: for the web and beyond.** 4. ed. Canadá: O'reilly Media Inc., 2015.

SAKAI, Hiroshi; NAKATA, Michinori; WATADA, Junzo. Nis-Apriori-based rule Generation with three-way decisions and its application systems in SQL. **Information Sciences,** v. 507, p. 755-771, jan. 2018. Disponível em: https://doi.org/10.1016/j.ins.2018.09.008. Acesso em: 28 ago. 2024.

SCHUNEIDER, Luís Felipe. **Mineração de Dados** - Conceitos. Porto alegre: Universidade Federal do Rio Grande do Sul, 2002.

SOTARO, Katsumata *et al.* Website Classification Using Latent Dirichlet Allocation and Its Application for Internet Advertising *In*: INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS (ICDMW), 16., 2016. **Anais** [...] Espanha: IEEE, 2016. Disponível em: https://doi.org/10.1109/ICDMW.2016.0083. Acesso em: 28 ago. 2024.

SUMITHRA, R.; PAUL, Sujni. Using distributed apriori association rule and classical apriori mining algorithms for grid-based knowledge discovery. *In*: INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION, AND NETWORKING TECHNOLOGIES, 2., 2010, India. **Anais** […]. India: IEEE, 2010. Disponível em: https://ieeexplore.ieee.org/document/5591577. Acesso em: 28 ago. 2024

VICTORINO, M. C; PINHEIRO, M. S; SANTOS, R. F. Organização da informação e do conhecimento em sistemas de informação transacionais para o seu reuso em sistemas de apoio a decisão. *In*: BAPTISTA, D. M; ARAÚJO JÚNIOR., R. H. (org). **Organização da informação abordagens práticas**. Brasília: Thesaurus, 2015. p. 219-247.

WITTGENSTEIN, Ludwig. **Investigações filosóficas**. Traduação de Marcos G. Montagnoli. 9. ed. São Paulo: Editora Vozes, 2014.