Identificação de Produtos em Descrições Textuais de Compras: Uma Proposta para Portais de Transparência Pública

Resumo. *Os portais de transparência vêm se constituindo em importantes canais de comunicação entre o governo e a sociedade. No entanto, nem sempre o formato das informações apresentadas é o mais apropriado. Por exemplo, as descrições de compras em formato de texto dificultam a análise dessas compras para a identificação do produto adquirido, e a posterior comparação entre as compras. O grande volume de dados inviabiliza uma identificação manual. Dessa forma, o objetivo desse trabalho é identificar automaticamente os produtos que são especificados de forma textual nas descrições de compras. Para isso, é proposto um processo de descoberta de conhecimento em dados textuais capaz de gerar regras que possibilitam a identificação de produtos a partir das descrições textuais de compras.*

Palavras-chave: *transparência pública, mineração de texto, tratamento de dados, processamento intensivo de dados, big data*

Abstract. *Transparency portals are becoming an important communication channel between government and society. However, the format of the information made available in these portals is not always the most appropriate. For example, descriptions of purchases in text format make it more difficult to analyze these purchases and later compare them with other purchases. Due to the large volume of data, manual identification is unfeasible. Thus, the objective of this work is to automatically identify products which are specified in text form in descriptions of purchases. For this purpose, a knowledge discovery process for text data is proposed which can generate rules enabling the identification of products based on text descriptions of purchases.*

keywords: *public transparency, text mining, data processing, intensive data processing, big data.*

1. INTRODUÇÃO

Com o avanço da Internet e das tecnologias digitais, questões de transparência eletrônica, democracia digital, governo aberto, ciberdemocracia e outros termos que associam a atuação governamental às ferramentas apoiadas pelo uso de Tecnologia da Informação vêm ganhando cada vez mais importância para a sociedade. Os ecossistemas digitais são ambientes que estimulam a participação cidadã e consequentemente aumentam o grau de democratização dos governos.

Acompanhando essa tendência, a área de transparência tem encontrado um terreno fértil para promover o estimulo ao controle social dos gastos públicos. Os portais de transparência pública vêm se transformando em importantes canais de comunicação entre o governo e a sociedade. Por meio desses portais, o cidadão tem acesso a uma série de informações que facilitam o acompanhamento e o controle das atividades governamentais.

Visando atender à crescente demanda por informações públicas, o governo brasileiro tem se empenhado para disponibilizar seus dados, tendo inclusive criado legislações específicas, Lei

Complementar 131(BRASIL, 2009), para garantir a disponibilização de dados governamentais. Porém, a simples disponibilização de dados na Internet não garante o aumento do grau de transparência governamental. Isso acontece porque a maioria dos dados disponibilizados para o cidadão não foram concebidos com esse propósito. Em geral, as informações são oriundas de sistemas coorporativos cujo objetivo é propiciar o controle administrativo das contas públicas e por isso nem sempre o seu formato é o mais apropriado para o cidadão entender o que realmente elas representam.

Dentre essas informações não tratadas, estão as descrições de compras feitas pela Administração Pública. Os produtos comprados são descritos em formato textual de livre preenchimento, o que inviabiliza a comparação entre as compras similares e prejudica o acompanhamento sistemático dos gastos.

Outro agravante nesse contexto é o elevado volume de dados disponibilizados diariamente por esses sites. Apesar da grande quantidade de informações apresentadas permitir uma maior abrangência e mais insumo para que o cidadão possa acompanhar a atuação governamental, a falta de mecanismos de classificação e organização dessas informações (com relação a importância desses gastos) acaba fazendo com que dados relevantes fiquem escondidos no grande volume de informações disponibilizadas, dificultando o entendimento, a comparação e reuso desses dados.

Assim, a questão de pesquisa que esse trabalho aborda é: como identificar de forma automatizada os produtos a partir das especificações textuais que são usadas para caracterizá-los nas descrições dos gastos que são apresentados nos portais de transparência pública?

Logo, o objetivo dessa monografia é fazer a identificação dos produtos mais comprados pela Administração Pública, por meio da análise das descrições textuais de compras, apresentadas nos portais de transparência. Sendo assim, o problema a ser tratado consiste em identificar de forma automatizada os produtos a partir das especificações textuais que são usadas para caracterizá-los nas descrições dos gastos que são apresentados nos portais de transparência pública.

Para isso, considerando-se uma frase como sendo uma sequência de tokens[1] contínuos e partindo-se da premissa de que descrições de produtos similares apresentam alguma sequência de tokens iguais, considerou-se a seguinte hipótese: “se forem identificadas as sequências de tokens que mais se repetem em um determinado conjunto de descrição de compras, então, essas sequências de tokens caracterizarão os produtos mais comprados desse conjunto de descrições”.

Optou-se pela caracterização apenas dos produtos mais comprados devido à grande quantidade de diferentes produtos que a Administração Pública pode adquirir (centenas de milhares), visto que, quanto mais comprado for um determinado produto, mais informações de compras se tem desse produto, tornando-se mais fácil identificar-se padrões de compras e consequentemente possíveis fraudes. Logo, esse trabalho pretende criar regras de identificação para os produtos mais comprados, otimizando assim o esforço de identificação de tais produtos.

Para a avaliação, desenvolveu-se uma metodologia capaz de analisar os resultados obtidos pela aplicação da solução proposta, sendo que, essa metodologia de avaliação permite verificar a qualidade dos resultados obtidos pelos diversos experimentos executados durante o processo de validação.

Dessa forma, essa monografia apresenta como contribuição principal a proposta de um método capaz de gerar regras de identificação de produtos a partir de descrições textuais de compras, porém, outras contribuições intermediárias também resultam dessa pesquisa, como a

[1] Em mineração de texto, define-se token como sendo a unidade mínima de um texto. No contexto dessa monografia, considera-se token como sendo sinônimo de palavra.

proposta de um algoritmo de geração de frases, de um algoritmo de poda de sub frases e o desenvolvimento de uma metodologia de avaliação dos resultados.

O restante desse trabalho está organizado da seguinte forma: A Seção 2 faz uma revisão de alguns trabalhos relacionados e as Seções 3 e 4 apresentam a proposta e a validação dessa proposta respectivamente. Na Seção 5 são apresentadas algumas possíveis aplicações para o método desenvolvido. Finalmente, na Seção 6 é feita a conclusão do artigo.

2. TRABALHOS RELACIONADOS

Estudando-se a literatura relacionada a portais de transparência pública, Hong (2014) faz uma análise de portais de governo aberto, na perspectiva da transparência para accountability. Este estudo tem como objetivo avaliar se a atual estrutura e organização de alguns dos portais de governo aberto é adequada para apoiar a transparência na prestação de contas. Sendo assim, Hong (2014) estabelece um conjunto de requisitos como base de características-chave da divulgação de dados sobre governo aberto e avaliação de transparência. Os resultados dos estudos sugeriram que este tipo de portal não possui elementos organizacionais necessários para apoiar plenamente os cidadãos comuns envolvidos em esforços de responsabilização pública. No entanto, a criação desses elementos organizacionais muitas vezes passa pela necessidade de estruturação de dados que estão originalmente expressos em formato de texto.

Atualmente, já existe uma série de trabalhos que se propõem a extrair informações relevantes de dados textuais gerados pela Administração Pública. Nesse sentido, CARVALHO et al. (2013) e CARVALHO et al. (2014b) sugerem uma metodologia para a formulação de um banco de preço da Administração Pública Federal Brasileira a partir dos dados de compras que são apresentados no Portal da Transparência do Governo Federal Brasileiro. Essas compras vêm descritas em formato textual e carecem do emprego de técnicas de mineração de texto para se extrair o produto correspondente a cada uma das descrições de compras.

A abordagem proposta está dividida em 6 passos. Primeiro, são selecionadas, do banco de dados do portal, todas as notas de empenho referentes a um determinado período. Depois, para cada uma dessas notas de empenho são recuperados os códigos de material das compras descritas. O passo seguinte é a filtragem do conjunto de dados referente a um código de material específico. No quarto passo, utiliza-se esses resultados da filtragem e emprega-se um novo filtro, baseado na utilização de palavras chave, a fim de se determinar um produto específico. Posteriormente, filtra-se o conjunto de dados resultante por faixa de preços, e finalmente calcula-se o preço de referência para o produto em questão.

O primeiro e o segundo passo são executados com o auxílio de uma ferramenta de ETL (Extract, Transform, Load). O terceiro passo (a filtragem pelo código de material) é executado através de consultas SQL (Structured Query Language), diretamente no banco de dados. Na filtragem por palavras chaves (quarto passo), especialistas definem quais palavras devem estar contidas e quais palavras não podem estar presentes na descrição de uma determinada compra, para que um determinado produto possa ser caracterizado. Isso permite a identificação dos produtos.

No entanto, mesmo após a caracterização do produto, ainda há uma grande variabilidade na faixa de preço paga. Essas diferenças, em muitas situações decorrem das diferentes formas de se quantificar um produto (por exemplo, diferentes unidades de medidas). Sendo assim, durante o passo 5 são aplicadas técnicas de clusterização, para cada grupo de produtos identificados, considerando-se que produtos quantificados de forma igual ficam em um mesmo cluster. Ainda nesse passo, os especialistas definem rótulos para cada um dos clusters gerados, sendo que um

produto será totalmente caracterizado a partir da combinação entre o nome do produto (identificado a partir da combinação de palavras chaves) com o rótulo definido pelos especialistas. Esses rótulos são escolhidos em uma lista que traz as palavras com maior probabilidade de definir um determinado cluster. Finalmente, após a qualificação dos produtos, utiliza-se os preços pagos por tais produtos a fim de se calcular uma faixa de preço de referência para esse produto. Nessa abordagem, especialistas precisam definir qual conjunto de palavras deve ser utilizado para caracterizar cada um dos produtos definidos como identificáveis. A definição de quais produtos irão compor o banco de preços também é feita pelos especialistas.

CARVALHO et al. (2014a) usam redes bayesianas (Friedman, Geiger, & Goldszmidt, 1997) para identificar e prevenir o fracionamento de compras, uma espécie de fraude utilizada para burlar o processo licitatório exigido por lei. No Brasil, compras inferiores a um determinado valor (R$ 8.000,00) são dispensadas do procedimento licitatório. No entanto, uma fraude comum, para enquadrar compras de valores superiores nesse tipo de dispensa é o fracionamento de uma mesma compra em várias outras de valores inferiores ao limite definido por lei. O objetivo principal desse trabalho é tentar identificar as compras consideradas suspeitas de terem sido fracionadas, a fim de permitir que providencias possam ser tomadas antes da consumação de um gasto irregular.

Essa identificação de compras suspeitas é feita através do uso de redes bayesianas e utiliza uma série de atributos estruturados durante o processo de classificação. No entanto, também se faz necessária a identificação dos produtos que estão sendo especificados de forma textual nos editais de compra.

MARZAGÃO (2015) apresenta uma outra abordagem para o problema de identificação de produtos e serviços que são adquiridos pela Administração Pública. Esse trabalho utiliza um cadastro de materiais e serviços adotados pelo Governo Federal Brasileiro no sistema SIASG (Sistema Integrado de Administração de Serviços Gerais) como dado de treinamento, e a partir deste cadastro, tenta classificar as compras utilizando o algoritmo de Máquina de Vetor de Suporte (CORTES; VAPNIK, 1995). Essa abordagem atingiu uma acurácia de 83,35%, e segundo o autor os erros encontrados foram ocasionados por duas causas principais: falhas no conjunto de dados de treinamento e problemas de frequência de classes, pois algumas classes de produtos, por não serem compradas frequentemente, não forneciam informações suficiente para o algoritmo de aprendizado de máquinas.

Visando atender a necessidade de processamento requerida pelo grande volume de informações que compõe as bases de dados de compras governamentais, PAIVA e REVOREDO (2016) apresentaram uma solução escalável para o problema de identificação de produtos em descrições textuais de compras. PAIVA e REVOREDO (2016) propuseram um modelo de identificação de produtos baseado em palavras chaves, semelhante ao processo utilizado em (CARVALHO et al., 2013) e (CARVALHO et al., 2014a). No entanto, na abordagem sugerida em (PAIVA e REVOREDO; 2016), ao invés de se empregar ferramentas de ETL e processamento sequencial, foi desenvolvido uma arquitetura que possibilita o processamento paralelo, resolvendo questões ligadas a limitações na capacidade de processamento.

O foco principal desse trabalho foi a proposta de uma arquitetura de processamento baseado no paradigma de programação MapReduce (DEAN; GHEMAWAT, 2008) e no framework hadoop (WHITE, 2012), que roda em clusters de computadores. Dessa forma, aumentos expressivos na quantidade de registros textuais a serem analisados e identificados podem ser compensados pela inclusão de novos computadores ao cluster utilizado.

Saindo do contexto da Administração Pública, mas ainda dentro do desafio de se extrair informações de dados textuais, algumas iniciativas têm se destacado no sentido de utilizar o

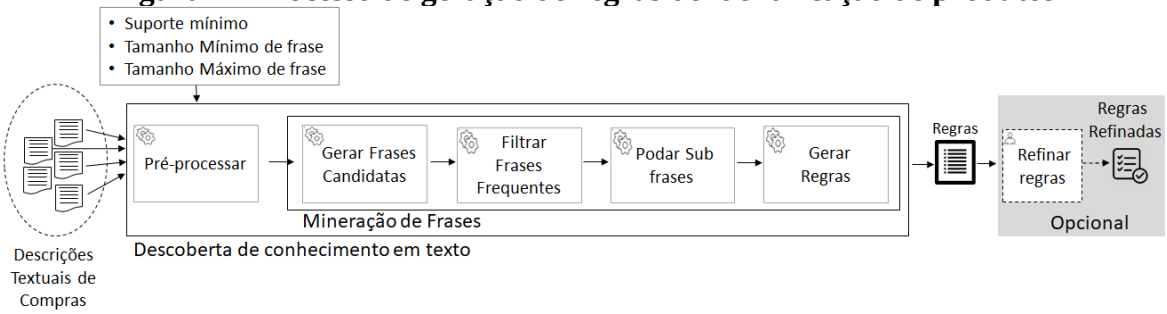
modelo de Bag of Phrases[2], em um contraponto ao tradicional Bag of Words (Salton, Wong, & Yang, 1975). Dentre essas iniciativas estão (REN et al., 2015), (LIU et al., 2015) e (EL-KISHKY et al., 2014). Essas abordagens, ao invés de trabalharem com os tokens de forma individualizada, consideram sequências de tokens, que formam frases, a fim de agregar mais expressividades as variáveis tratadas.

Pela análise dos trabalhos relacionados, verificou-se que esses apresentavam algumas limitações, dentre as quais pode-se destacar: a necessidade de utilização de um conjunto de dados de treinamento, o que normalmente não está disponível, ou a necessidade de definição de palavras chaves, por parte de especialistas, para se realizar a extração de informações úteis de conjuntos de dados textuais. Logo, a principal contribuição desse artigo em relação aos demais trabalhos que também se propõem a extrair informações de dados textuais governamentais é a proposta de uma técnica de extração de conhecimento baseada no modelo Bag of Phrases, capaz de minerar as frases que melhor representam o conteúdo de um determinado texto e que não exige um conjunto de dados previamente rotulados e nem tem a necessidade de intervenção de especialistas durante o processo de descoberta de conhecimento.

3. PROPOSTA

Essa pesquisa aborda o problema científico da identificação automática dos produtos adquiridos em uma compra a partir da sua descrição textual. Para isso é proposto um método que recebe um conjunto de descrições textuais e retorna um conjunto de regras de identificação. Essas regras posteriormente são utilizadas para identificar os produtos adquiridos a partir das suas descrições textuais. O método proposto está ilustrado na Figura 1. Ele está dividido em cinco passos obrigatórios e um opcional, sendo que essa etapa adicional é executada dependendo da disponibilidade de especialistas do domínio. O objetivo desse sexto passo opcional é melhorar a forma de representação do conhecimento das regras criadas, assim como fazer algumas adaptações nas regras, de modo que essas possam ser mais adequadas para os propósitos finais da classificação gerada, bem como para melhorar os resultados obtidos.

Figura 1 - Processo de geração de regras de identificação de produtos



Fonte: Elaboração do Autor

No contexto desse trabalho, uma frase é definida como uma sequência contigua de tokens. Sendo assim, nessa monografia, a tarefa de mineração de frases pode ser caracterizada pela agregação e contagem de todas as sequências iguais de tokens contíguos que satisfaçam a um

[2] Bag of Phrase: modelo de representação utilizado no tratamento de dados textuais. Nesse modelo o texto é representado pela contagem das frases que o compõem, ignorando-se a gramática e a ordem das frases.

suporte mínimo. Ou seja, a mineração de frases se propõe a identificar os padrões sequenciais de tokens que mais se repetem em um determinado conjunto de dados textuais.

Dessa forma, as seguintes propriedades, definidas em (EL-KISHKY et al. 2014) e (LIU et al. 2015), deverão ser atendidas no processo de mineração de frases:

- Frequência: A qualidade mais importante quando se julga se uma frase retransmite informações relevantes sobre um tópico é a sua frequência de utilização dentro do tópico. Uma frase que não é frequente dentro de um tópico, provavelmente não é importante para esse tópico.
- Completude: Se uma frase longa satisfaz ao critério da frequência, então, as sub frases dessa frase longa também irão satisfazer a este critério, porém, serão menos informativas do que a frase mais longa, e dessa forma não precisam ser consideradas na mineração, pois a frase mais longa é mais completa.

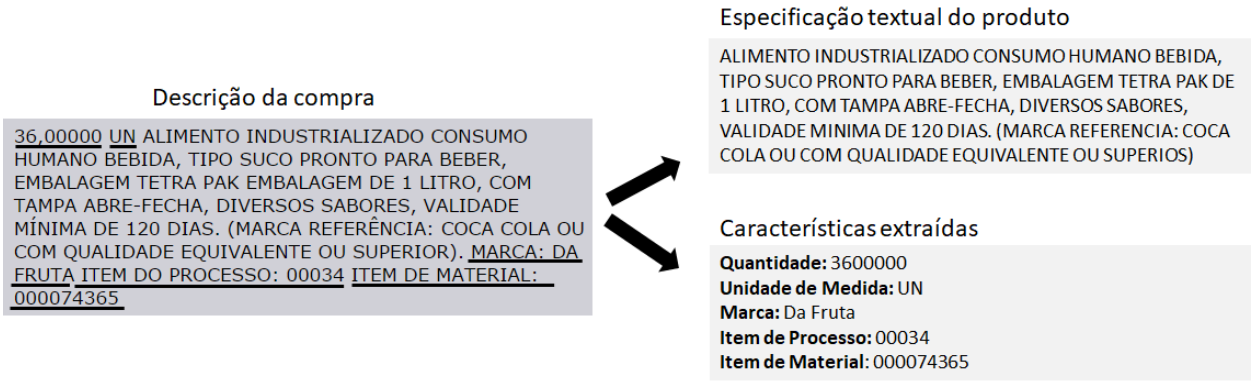
Devido as características dos dados de portais de transparência, grandes volumes de informações com cargas diárias e incrementais. A solução proposta deve ser capaz de processar quantidades massivas de dados. Para atender a esse requisito, todo o processo foi concebido para rodar utilizando o Apache Spark (ZAHARIA et al. 2010), um framework para processamento de grandes volumes de dados (Big Data) que roda de forma paralela em cluster de computadores.

As seções seguintes descrevem em mais detalhes as etapas do método proposto.

3.1. PRÉ-PROCESSAMENTO

O pré-processamento é a primeira etapa do método, e tem o objetivo de preparar o conjunto de dados para as atividades subsequentes. Essa etapa de pré-processamento retira informações que estão presentes no campo de descrição da compra, mas que não fazem parte da especificação textual do produto. O principal objetivo desse procedimento é a eliminação de informações desnecessárias que possam prejudicar a análise das sequências de palavras geradas.

Figura 2 - Resultado do Pré-processamento



Fonte: Elaboração do Autor

Na Figura 2 é ilustrado o resultado do pré-processamento de uma descrição de compra. Nesse procedimento, algumas informações são identificadas e extraídas, através das técnicas

enunciadas em (ETZIONI et al. 2005). Essas técnicas pregam a utilização de templates na atividade de extração de informações de dados textuais. Para isso, cada template é utilizado para extrair um tipo de relação específica entre as palavras que aparecem no texto. Por exemplo, o template “tais como” na frase, “Cidades tais como Rio de Janeiro e São Paulo” permite-nos concluir que os termos Rio de Janeiro e São Paulo são instâncias do conceito cidade.

Para o caso das descrições textuais das compras, formam identificados templates específicos para esse contexto, e as relações obtidas pela aplicação dos templates ocorrem entre a compra em si e o termo referenciado pelo padrão buscado. Essas relações permanecem verdadeiras nas diferentes descrições de compras porque parte dessas descrições é gerada de forma automatizada, a partir de alguns dados estruturados, enquanto que, a especificação do produto, propriamente dito, é feita manualmente. Dessa forma, o termo: “Marca: Da Fruta” evidencia que a compra que está sendo descrita se refere a um produto cuja marca é Da Fruta.

Sendo assim, a utilização de simples templates de identificação permite a extração de uma série de características da compra que, após a devida classificação do produto, ao final de todo o processo, podem agregar maior conhecimento a respeito das informações apresentadas. Portanto, apesar dessas características extraídas não serem utilizadas na proposta apresentada, elas são utilizadas no processo de validação dessa proposta, assim como para as aplicações que podem ser realizadas com os resultados obtidos com o emprego da técnica desenvolvida. Logo, além da retirada de termos que possam prejudicar a mineração textual, essa fase também é responsável pela extração de algumas características da compra, que ao final do processo permitem agregar maior conhecimento a respeito dos produtos que estão sendo adquiridos.

Cabe ressaltar que esse procedimento não serve para a identificação do produto que está sendo especificado na descrição da compra. Essa impossibilidade se dá porque não há um padrão na especificação textual dos produtos, uma vez que, essa especificação é feita de forma manual, e cada pessoa preenche a designação do produto de uma forma diferente.

Outra atividade realizada na etapa do pré-processamento é a filtragem dos dados que não se referem a compra de materiais, visto que, existem outros tipos de gastos que não dizem respeito à compra de produtos, como por exemplo, contratação de serviços e pagamento de pessoal. Essa pesquisa não considera os contratos de prestação de serviços pelo fato desses apresentarem grande variabilidade de características, o que faz com que cada contratação seja única.

Durante o pré-processamento, também é realizado um tratamento no texto de forma que todas as letras presentes nas descrições de compras sejam passadas para o formato de letra maiúscula e que todos os sinais de acentuação sejam retirados.

A saída dessa etapa é o conjunto pré-processado das descrições textuais dos produtos. A próxima etapa, descrita na seção seguinte, tem por objetivo encontrar frases candidatas a identificação de um produto.

3.2. GERAÇÃO DE FRASES CANDIDATAS

Apesar do método proposto apresentar um enfoque estatístico, com o intuito de se diminuir o conjunto de possíveis combinações de palavras, assim como para manter a expressividade das frases geradas, algumas considerações semânticas foram feitas:

- Uma frase só pode ser formada se ela estiver contida dentro de uma determinada sentença. Nessa pesquisa, considera-se sentença como sendo uma sequência de palavras delimitada por sinais de pontuação que determinam o final de um período (ponto final, ponto de exclamação ou ponto de interrogação).

- Se um determinado token W está localizado na posição n de uma sequência de tokens de uma sentença, para que esse token W faça parte de uma frase, é necessário que todos os demais tokens localizados nas $(n-1)$ posições anteriores da sequência, também façam parte dessa frase. Essa restrição foi formulada para garantir maior grau de expressividade para as frases formadas, visto que, na língua portuguesa o significado de uma frase vai se completando da esquerda para a direita.

O Algoritmo 1 faz uso dessas considerações e realiza a geração de frases a partir de um conjunto de especificações textuais de compras.

Algoritmo 1 - Algoritmo de geração de Frases Candidatas

Algoritmo 1: Geração de Frases Candidatas

Entrada:
Conjunto de Especificações de Produtos *E*,
tamanho mínimo da frase *min* e
tamanho máximo da frase *max*

Saída:
Vetor com frases construídas

```
1. Início
2.     frases=[]
3.     Para cada especificação e em E Faça:
4.         Sentenças= SeparaSentenças(e)
5.         Para cada sentença em sentenças Faça:
6.             Para m entre (min,max) Faça:
7.                 SE (tamanho(sentença)>=m)
8.                     frases.insere(sentença[0:m])
9.                 Fim
10.            Fim
11.        Fim
12.    Fim
13.    retorna frases
14. Fim
```

Fonte: Elaboração do Autor

O algoritmo de geração de frases candidatas recebe como entrada um conjunto de especificações textuais de produtos e como parâmetro um tamanho mínimo e outro máximo para as frases a serem geradas, sendo que, o tamanho de uma frase é medido pelo número de palavras que compõem essa frase. A saída do algoritmo proposto será o conjunto de todas as frases geradas.

Cabe ressaltar que os primeiros trabalhos que abordaram a questão de mineração de frases (REN et al., 2015), (LIU et al., 2015) e (EL-KISHKY et al., 2014) já propunham a utilização de algoritmos de geração de frases, porém, nessa monografia foi proposto um algoritmo específico para os propósitos dessa pesquisa, que considera peculiaridades semânticas da língua portuguesa.

Para exemplificar o funcionamento do algoritmo de geração de frases candidatas, na Figura 3 é apresentada a especificação textual de uma determinada compra, que seria um elemento pertencente ao conjunto de especificações de produtos *E*, que funciona como entrada para o

algoritmo, e as frases resultantes da aplicação desse algoritmo para o caso de se utilizar os parâmetros de tamanhos mínimo e máximo de frases como sendo 5 e 10, respectivamente. Nesse caso de exemplo, a especificação de entrada é composta por duas sentenças, que são delimitadas por um ponto final, e a saída é composta pelo conjunto de frases geradas a partir dessas duas sentenças.

Figura 3 – Exemplo do processo de Geração de Frases

Fonte: Elaboração do Autor

3.3. FILTRAGEM DE FRASES FREQUENTES

Após a geração das frases candidatas, o passo seguinte é a agregação das frases iguais, a fim de se contar o número de ocorrências de cada uma das frases geradas. Sendo assim, para cada frase gerada pelo algoritmo de geração de frases candidatas é feita uma verificação e contagem de todas as frases coincidentes. O algoritmo executado nessa etapa não é apresentado pelo fato de ser bem simples, uma vez que, ele apenas faz uma agregação e contagem das frases iguais e desconsidera aquelas frases cuja contagem não atinja a um determinado suporte mínimo, passado como parâmetro. Logo, ele recebe como entrada um conjunto de frases geradas (saída da etapa anterior), conta o número de ocorrências de cada uma dessas frases, e apresenta como saída o conjunto de frases cujo número de ocorrências tenha superado o suporte mínimo.

Dessa forma, cada frase estará associada a um número de ocorrências, e aquelas frases que tiverem esse número de ocorrências superior a um suporte mínimo, passado como parâmetro, prosseguem no processamento, enquanto que, as frases cujo número de ocorrências for inferior a esse suporte são desconsideradas.

Essa etapa tem o objetivo de atender ao critério da frequência, e o seu funcionamento é ilustrado pela Figura 4, sendo que, nesse exemplo considera-se o suporte mínimo de 30. Dessa forma, na parte (a) da Figura 4 é mostrado um conjunto de frases geradas, que foram obtidas pelo algoritmo de geração de frases e são a entrada do algoritmo de filtragem de frases. Na parte (b) são mostradas algumas dessas frases já grupadas e com as respectivas quantidades de ocorrências de cada uma dessas frases. Por fim, na parte (c) da Figura 4 é apresentada a saída do algoritmo para o conjunto de frases e suportes considerados.

Figura 4 – Exemplo do processo de filtragem de frases

Fonte: Elaboração do Autor

3.4. PODA DE SUBFRASES

EL-KISHKY et al. (2014) definem duas propriedades na mineração de frases:

- Lema do fechamento para baixo: Se uma frase G não é frequente, então as super frases de G (frases que contêm G) também não serão.
- Antimonotonicidade dos dados: Se um documento não contém frases frequentes de comprimento n, o documento não contém frases frequentes de comprimento maior que n.

A aplicação dessas propriedades ao conjunto de frases resultante do passo anterior serve para reduzir a quantidade das frases decorrentes do processo de mineração. Sendo assim, se uma frase G, formada pela sequência de palavras $w_1\ w_2\ \ldots\ w_n$ atende ao requisito do suporte mínimo, então, todas as suas sub frases $G' = w_1\ w_2\ \ldots\ w_k$, com $k\ < n$, também atenderão a esse suporte, porém, elas não precisarão ser analisadas, uma vez que as frases maiores (em que elas estão contidas) já contemplam ao requisito necessário (suporte mínimo). Logo, é executado uma poda aplicando essa propriedade de forma a reduzir o número de frases mineradas.

No Algoritmo 2 é mostrado o processo de poda das sub frases. Esse algoritmo recebe como entrada todas as frases geradas que atenderam ao critério do suporte mínimo, e oferece como saída apenas as super frases (ou seja, frases contidas em outras frases maiores que também atendam ao

requisito do suporte mínimo são desconsideradas). Essa etapa tem o objetivo de atender ao critério da completude (definido no início da seção 3).

Algoritmo 2 - Algoritmos de Poda de Sub Frases

Algoritmo 2: Poda Sub Frases

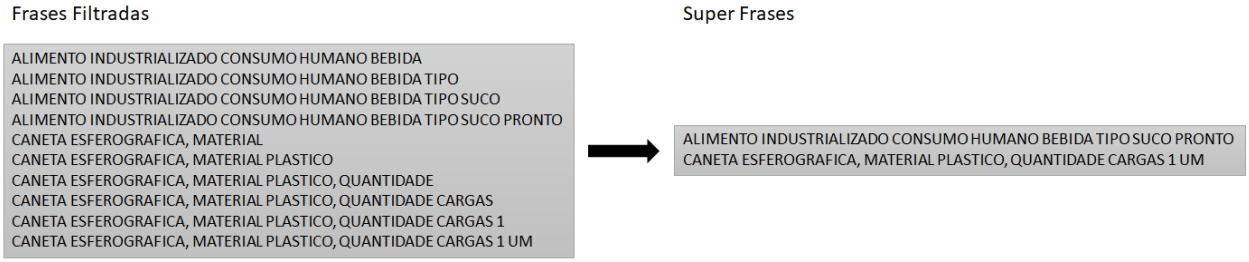
Entrada:
Vetor H, com as frases que satisfazem o suporte mínimo
Saída:
Vetor com SuperFrases

```
1.  Início
2.      frasesDeQualidade=[]
3.      Ordena(H) // ordena em ordem decrescente de tamanho
4.      Para cada frase h em H Faça:
5.          SuperFrase=Verdadeiro
6.          Para cada frase sp em frasesDeQualidade Faça:
7.              Se h em sp Então:
8.                  SuperFrase=Falso
9.                  continua
10.             Fim
11.         Fim
12.         Se SuperFrase=Verdadeiros Então
13.             frasesDeQualidade.insere(h)
14.         Fim
15.     Fim
16.     retorna frasesDeQualidade
17. Fim
```

Fonte: Elaboração do Autor

Para exemplificar o funcionamento do algoritmo de poda de sub frases, a Figura 5 apresenta, no lado esquerdo, um conjunto de frases resultantes do processo de filtragem de frases, ou seja, frases que tenham atendido ao suporte mínimo passado como parâmetro. Já o lado direito da Figura 5 representa a saída do algoritmo, considerando-se como entrada as frases apresentadas do lado esquerdo da figura.

Figura 5 - Exemplo do processo de poda de sub frases



Fonte: Elaboração do Autor

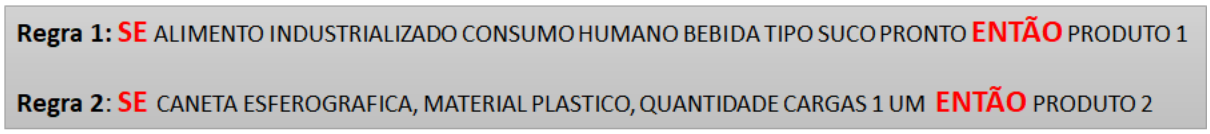
3.5. GERAÇÃO DE REGRAS

A última etapa do processo é a geração das regras de identificação. As regras são do tipo: “SE antecedente ENTÃO consequente”, sendo o antecedente a premissa, definida por uma determinada frase, e o consequente o produto a ser identificado a partir da premissa.

Logo, cada frase resultante do processo de poda de sub frases dá origem a uma regra distinta. Dessa forma, assume-se que todas as compras que se enquadrarem em uma determinada regra de identificação (ou seja, todas as compras cuja especificação tenha alguma frase que coincida com uma determinada frase considerada como antecedente, que resultou do processo de mineração de frases) se referem a um mesmo tipo de produto.

Portanto, a regra 1 vai identificar um determinado produto 1, a regra 2 identifica um determinado produto 2 e assim por diante. Na Figura 6 são apresentados dois exemplos de regras geradas a partir das super frases resultantes da etapa de Poda de Subfrases.

Figura 6 – Regra Gerada



Fonte: Elaboração do Autor

3.6. REFINAMENTO DE REGRAS

Conforme dito anteriormente, a solução proposta ainda prevê uma sexta etapa. Porém, a etapa de refinamento de regras é opcional e depende da disponibilidade de especialista de domínio para a realização dessa tarefa, visto que, essa última etapa carece de uma interação humana. Essa etapa tem o objetivo de melhorar a forma de representação do conhecimento expressa pelos consequentes das regras geradas, bem como possibilitar a agregação, ou eliminação de regras de acordo com o grau de especificidade, ou generalidade, que se deseja dar no processo de identificação das compras.

Durante essa etapa, os especialistas analisam as regras geradas e fazem a seleção e validação dessas regras, bem como a escolha de consequentes semanticamente mais apropriados para cada regra. Logo, o esforço dos especialistas nessa fase se resume em fazer a seleção e validação dos antecedentes e reformular os consequentes de cada regra.

- Seleção/validação dos antecedentes: Esse procedimento tem duas finalidades, a primeira se dá porque, apesar das frases tenderem a ter um alto grau de expressividade, pois, elas atingiram uma frequência alta de ocorrência, em algumas situações elas podem não transmitir informações capazes de discriminar um determinado produto. Outra razão que justifica o benefício da interação humana é a definição do grau de especificidade que se deseja dar a um determinado produto. Por exemplo, um produto pode ser identificado como suco de laranja ou simplesmente como suco, dependendo da análise que se deseja fazer, e a seleção dos antecedentes das regras de identificação tem importante papel nesse processo.

- Reformulação de consequentes: Um papel relevante, executado por especialistas, é a interpretação dos antecedentes das regras, a fim de definir consequentes

semanticamente mais apropriados. Por exemplo, nas regras apresentadas na Figura 6, pode-se substituir os consequentes PRODUTO 1 e PRODUTO 2 por SUCO INDUSTRIALIZADO e por CANETA ESFEROGRÁFICA nas regras 1 e 2, respectivamente. Outra vantagem dessa atividade é a possibilidade de se definir consequentes iguais para regras diferentes, mas que tenham o mesmo conteúdo informacional. Por exemplo, um especialista pode definir um mesmo rótulo para os antecedentes "dipirona, solução oral 500 mg/ml" e "novalgina gotas 500 mg/ml", associação essa que seria difícil de se fazer de forma automatizada.

4. AVALIAÇÃO

Como citado anteriormente, a questão de pesquisa desse trabalho é "como identificar de forma automatizada os produtos a partir das especificações textuais que são usadas para caracterizá-los nas descrições dos gastos que são apresentados nos portais de transparência pública?", dessa forma, a avaliação foi desenvolvida de forma a verificar se a solução proposta foi capaz de responder a essa questão de pesquisa, e se a hipótese, "se forem identificadas as sequências de tokens que mais se repetem em um determinado conjunto de descrição de compras, então, essas sequências de tokens caracterizarão os produtos mais comprados desse conjunto de descrições", é verdadeira ou não. Sendo assim, essa seção apresenta um estudo que tem o objetivo de avaliar os resultados obtidos pela aplicação da solução proposta em um conjunto de descrições de compras (itens de empenho referentes a compra de material) que são apresentadas no Portal da Transparência do Governo Federal. Esses dados estão disponíveis para download no referido portal.

4.1. PROJETO DE AVALIAÇÃO

O projeto de avaliação está dividido em duas partes: a primeira consiste da avaliação das regras geradas, enquanto que a segunda verifica a qualidade dos resultados obtidos no processo de identificação de compras propriamente dito, sendo que, em ambos os casos, a validação é feita após a aplicação das regras a um conjunto de descrições textuais de compras referentes ao ano de 2015 dos dados apresentados no Portal da Transparência do Governo Federal, sendo que, os dados referentes ao mês de janeiro foram utilizados para a geração das regras e os dados referentes aos outros meses do ano foram utilizados para a aplicação das regras criadas. Cabe ressaltar que se optou por utilizar-se os dados de 2015 nessa fase, para utilizar-se os dados dos anos seguintes (de 2016 a 2019) nas aplicações apresentadas na Seção 5.

4.1.1. Avaliação das Regras

A avaliação das regras geradas é feita através da aplicação de um método de clusterização. Dessa forma, para cada uma das regras de identificação, são utilizadas as compras que se enquadraram nessas regras. Em tal procedimento, utiliza-se os atributos da classificação da

natureza de despesa detalhada[3], como variáveis para o cálculo dos clusters formados pelas compras que se enquadram nas regras.

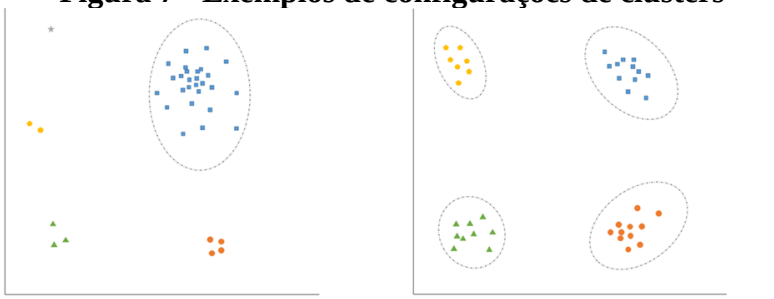
Partindo-se da premissa de que compras que se referem a um mesmo produto possuem a mesma classificação de natureza de despesa, considerou-se que em uma situação ideal, o processo de clusterização de uma regra perfeita iria gerar um único cluster com todas as compras identificadas por essa regra concentradas nesse cluster único, pois as compras identificadas de forma correta iriam corresponder a compras de um mesmo produto, e consequentemente possuiriam os mesmos atributos de natureza de despesa detalhada.

No entanto, numa situação real, outras variáveis externas interferem na avaliação do processo como um todo, como por exemplo a inserção de dados de natureza de despesa errada por parte dos usuários responsáveis por tal atividade. Dessa forma, formulou-se as seguintes considerações gerais para a análise dos clusters formados:

- Caso haja vários clusters com quantidades equivalentes de ocorrências de compras, provavelmente a regra não é boa, pois ela está pegando muitos produtos com classificação de natureza de despesa diferentes, e provavelmente está identificando produtos diferentes como sendo iguais.
- Em situações em que haja poucos clusters, sendo que apenas um cluster concentra a maioria das compras, e os demais possuem poucas ocorrências, provavelmente essa regra é boa, e as ocorrências dispersas (outros clusters) supostamente tenham sido fruto de classificações erradas da natureza de despesa, por parte da pessoa responsável por inserir essa informação no sistema.

A seguir é apresentado um exemplo do resultado do processo de clusterização das compras identificadas por uma regra boa e por uma regra ruim, nas figuras 7.a e 7.b, respectivamente. Deve ser observado que essas figuras representam situações hipotéticas, e têm apenas o objetivo de ilustrar as configurações resultantes do processo de clusterização que caracterizam uma regra como sendo boa ou ruim. Os parâmetros utilizados para determinar a qualidade das regras são especificados adiante, ainda nessa seção.

Figura 7 - Exemplos de configurações de clusters



Fonte: Elaboração do Autor

[3] A Natureza da Despesa Detalhada é uma classificação utilizada pela Contabilidade Pública, e é composta por um código formado por 8 dígitos numéricos, divididos em 5 grupos, com a seguinte configuração: X.Y.ZZ.MM.NN. Nessa codificação, o 1º dígito indica o código da categoria econômica, o 2º dígito indica o grupo de natureza de despesa, os 3º e 4º dígitos indicam a modalidade de aplicação, os 5º e 6º dígitos indicam o elemento de despesa e os 7º e 8º dígitos indicam o sub elemento de despesa.

Conforme pode ser observado na Figura 7, quanto mais homogênea for a configuração do cluster obtido pelo processo de clusterização das compras enquadradas em uma determinada regra, melhor será a qualidade dessa regra. Por outro lado, quanto mais disperso for a configuração desses clusters, pior será a qualidade dessa regra, pois provavelmente essa regra está identificando produtos diferentes, o que não é o comportamento esperado para as regras.

4.1.2. Avaliação da Identificação

Dada a inexistência de um conjunto de dados previamente rotulado, associada a inviabilidade de se analisar individualmente cada um dos casos, a avaliação dos resultados obtidos no processo de identificação de compras é feita de forma qualitativa. No entanto, o processo de escolha da amostra a ser analisada qualitativamente é feito por procedimentos quantitativos.

Então, a avaliação utiliza uma abordagem empírica e emprega técnicas quantitativas e qualitativas. As técnicas quantitativas são empregadas para a seleção de um conjunto amostral com maior probabilidade de ter sido classificado erroneamente. Já as técnicas qualitativas são utilizadas para analisar a amostra de dados selecionada.

A abordagem também combina características descritivas e explanatórias. A parte descritiva tem o objetivo de descrever as características dos dados relativos a cada um dos tipos de produtos identificados, a fim de possibilitar a seleção dos casos em que as características do produto diferem dos demais da mesma espécie. Já a parte explanatória tem o objetivo de analisar de forma mais detalhada esses casos selecionados, a fim de avaliar se a identificação do produto foi feita corretamente ou não.

Logo, as compras são grupadas por produtos, e para cada um dos grupos de produtos busca-se por outliers. Sendo assim, partindo-se da premissa que produtos iguais tendem a apresentar características similares, pode-se inferir que aqueles produtos cujos atributos sejam discrepantes em relação aos demais têm maior probabilidade de terem sido identificados de forma errônea. Dessa forma, pode-se selecionar uma amostra com casos mais propensos a terem sido classificados erradamente, otimizando assim a atividade dos especialistas responsáveis pela análise qualitativa desses resultados.

4.1.3. Estudo de Caso

A proposta de geração de regras de identificação, descrita na Seção 3, utiliza, além dos dados de entrada (um conjunto de descrições textuais de compras), três tipos de parâmetros: tamanho mínimo de frase, tamanho máximo de frases e suporte. Sendo assim, utilizou-se como fonte de dados de entrada, para a geração das regras de identificação, as descrições de compras dos itens de empenho, apresentados no Portal da Transparência do Governo Federal, referentes ao mês de janeiro do ano de 2015, e executou-se um estudo de caso composto por uma série de experimentos, sendo que, para cada um dos experimentos utilizou-se uma combinação diferente dos parâmetros requeridos pelo processo de geração de regras, ou seja, variou-se os valores de tamanhos máximo e mínimo de frase e o suporte requerido.

Para a verificação dos resultados obtidos pelo processo de identificação das descrições textuais de compras, utilizou-se os dados das descrições de compra dos itens de empenho, apresentados no Portal da Transparência do Governo Federal, referentes aos meses de fevereiro a dezembro do ano de 2015. Uma vez realizados os experimentos, com as diferentes configurações,

analisou-se os resultados obtidos, a fim de se verificar a consistência das regras geradas com cada uma das configurações de parâmetros.

4.2. EXPERIMENTOS

Durante a execução do estudo de caso foram realizados seis experimentos, conforme a configuração apresentada na Tabela 1. A intenção da variação dos parâmetros utilizados é possibilitar uma análise do comportamento da solução proposta com diferentes configurações, a fim de se identificar aquela mais apropriada para a situação apresentada.

Tabela 1 - Parâmetros dos experimentos realizados

Identificação	Tam Mínimo Frase	Tam Máximo Frase	Suporte
Experimento 1	10	15	10
Experimento 2	10	15	15
Experimento 3	10	15	30
Experimento 4	6	9	15
Experimento 5	6	9	30
Experimento 6	6	9	50

Fonte: Elaboração do Autor

4.2.1. Análise das Regras Geradas

Conforme apresentado na subseção 4.1.1, as regras geradas são avaliadas pela aplicação de métodos de clusterização. Dessa forma, utilizou-se o algoritmo "Density Based Spatial Clustering of Application with Noise" – DBSCAN (ESTER et al ,1996) - para executar os testes.

O DBSCAN (ESTER et al ,1996) é um método de clusterização não paramétrico baseado em densidades. A escolha desse método se deu pelo fato dele não exigir que sejam informados previamente o número de clusters a serem encontrados.

Dessa forma, para cada uma das regras de identificação, aplicou-se o método de clusterização DBSCAN em todas as compras identificadas por essa regra, utilizando-se os atributos de natureza de despesa detalhada como variáveis a serem consideradas nesse processo de clusterização.

Para auxiliar na geração dos clusters, foi utilizada a ferramenta RapidMiner[4], sendo que, todos os 6 experimentos executados foram considerados. Dessa forma, classificou-se as regras em três categorias diferentes, de acordo com a configuração dos clusters gerados. Sendo assim, uma regra pode ser classificada como sendo boa, regular ou ruim.

- Regra Boa

Para uma regra ser considerada como boa, é necessário que o processo de clusterização gere um cluster com mais de 90% das compras classificadas nessa regra.

Um exemplo de uma regra considerada como boa é a regra R_22 do experimento 6, cujo antecedente é formado pela seguinte sequência de tokens: "caneta esferografica, material plastico, quantidade cargas 1 un" e que concentra 95,38% das ocorrências em um mesmo cluster.

[4] https://rapidminer.com

- Regra Regular

Para que uma regra seja considerada como regular, é necessário que o processo de clusterização gere 2 clusters cuja soma das percentagens das ocorrências supere os 90%. Essa consideração é feita para se atender aos casos em que um mesmo produto pode ser enquadrado em duas classificações de natureza de despesa detalhada diferentes. Por exemplo, o produto "Reagente para diagnóstico clinico" que pode ser enquadrado tanto na natureza de despesa detalhada 3.3.90.30.09 quanto na 3.3.90.30.36, que correspondem a material farmacológico e material hospitalar, respectivamente. Para o experimento 6, uma regra classificada como regular foi a regra cujo antecedente era formado pela seguinte sequência de palavras: "reagente para diagnostico clinico, tipo conjunto completo para" que possui os dois clusters mais populosos com 54,38% e 42,50% das ocorrências.

- Regra Ruim

Sempre que uma regra não se enquadrar em uma das duas situações anteriores (regra boa ou regular), ela é considerada como ruim. Um exemplo de uma regra considerada como ruim é a regra R_1 do experimento 6, cujo o antecedente é formado pela sequência de palavras ", numero de referencia quimica cas[5]", pois essa sequência de palavras permite-nos inferir que o produto que está sendo especificado possui algum tipo de elemento químico em sua composição, porém, diferentes tipos de produtos podem ser enquadrados nessa regra. Tal regra gerou 40 clusters distintos e apresentou a seguinte distribuição de frequência de ocorrência para os 5 clusters mais populosos: 57,04%, 14,59%, 11,44%, 4,99% e 3,79%.

Os experimentos geraram diferentes quantidades de regras, sendo que o experimento 6 foi o que gerou menos regras, totalizando 25 regras distintas. Dessa forma, para se comparar a qualidade das regras geradas em cada um dos experimentos, considerou-se todas as 25 regras geradas para o experimento 6, e para os demais experimentos considerou-se as 25 regras que tenham tido mais compras classificadas

Na Tabela 2 são apresentados os clusters obtidos para as compras classificadas de acordo com as regras geradas na execução do experimento 6. Essa tabela apresenta uma linha para cada uma das 25 regras geradas, e colunas com a identificação das regras, número de clusters gerados por regra, porcentagem de itens de compra classificados nos 5 clusters mais populosos de cada regra e classificação da regra de acordo com os critérios citados acima (boa, regular ou ruim).

Conforme pode ser observado na tabela 2, o experimento 6 gerou 15 regras boas, 5 regras regulares e 5 regras ruins.

A tabela com os resultados obtidos nos processos de clusterização dos demais experimentos foram omitidas por uma questão de otimização de espaço, porém, na Figura 8 é apresentado um quadro resumo com o resultado obtido por cada um dos experimentos após a classificação das regras de acordo com os critérios de qualidade de regras definidos anteriormente.

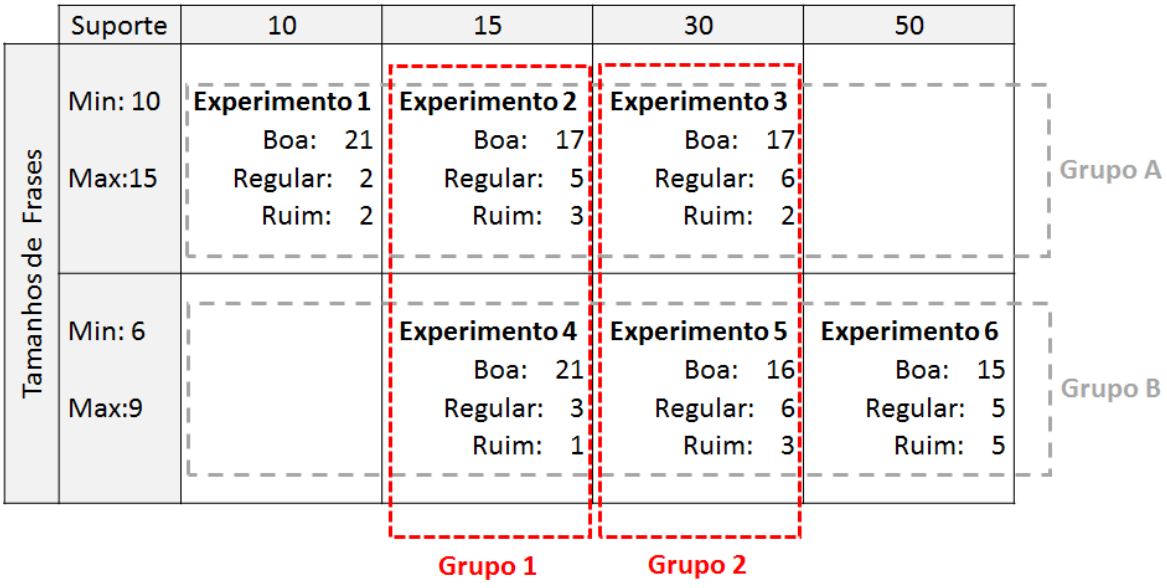
[5] O número de referência química CAS é um número de registro que todo os produtos químicos têm, esse número único de registro consta no banco de dados do Chemical Abstracts Service, uma divisão da Chemical American Society, e serve para identificar um determinado produto químico. Logo, todos os produtos químicos adquiridos pela Administração Pública Federal irão possuir a sequência de palavras "numero de referencia quimica cas", no entanto, tal sequência de palavras não é suficiente para identificar um determinado produto químico.

Tabela 2 - Clusters das regras gerados no experimento 6

Id Regra	Qtd de Clustres	Porcentagem de itens por clusters mais populosos (%)					Situação da Regra
		1º mais populoso	2º mais populoso	3º mais populoso	4º mais populoso	5º mais populoso	
R_1	40	57,04	14,59	11,44	4,99	3,79	Ruim
R_2	13	96,20	0,85	0,75	0,40	0,34	Boa
R_3	5	95,46	1,76	1,14	1,05	0,57	Boa
R_4	13	94,15	1,55	1,51	1,11	047	Boa
R_5	5	76,16	16,38	4,66	2,28	0,50	Regular
R_6	1	100,00	-	-	-	-	Boa
R_7	14	69,07	10,58	10,34	1,86	1,57	Ruim
R_8	12	79,45	6,37	3,54	3,14	2,08	Ruim
R_9	12	52,82	43,45	1,32	1,07	0,51	Regular
R_10	13	93,14	3,58	0,66	0,61	0,46	Boa
R_11	11	96,89	0,84	0,51	0,36	0,30	Boa
R_12	6	94,96	2,23	1,21	0,89	0,38	Boa
R_13	3	76,75	12,43	10,81	-	-	Ruim
R_14	7	93,41	2,61	1,62	1,06	0,67	Boa
R_15	14	97,41	0,76	0,27	0,26	0,22	Boa
R_16	5	94,28	1,87	1,58	1,48	0,78	Boa
R_17	17	53,17	14,36	13,45	8,44	2,66	Ruim
R_18	11	54,38	42,50	1,50	0,77	0,31	Regular
R_19	1	100,00	-	-	-	-	Boa
R_20	8	51,23	43,48	1,50	1,44	1,28	Regular
R_21	7	54,59	42,13	1,19	0,98	0,64	Regular
R_22	13	95,38	0,80	0,57	0,46	0,42	Boa
R_23	8	95,44	0,88	0,74	0,68	0,68	Boa
R_24	5	93,49	3,51	1,36	1,20	0,43	Boa
R_25	8	95,23	2,01	0,86	0,67	0,57	Boa

Fonte: Elaboração do Autor

Figura 8 - Qualidade das regras geradas por experimento



Fonte: Elaboração do Autor

Analisando-se separadamente os grupos de experimentos com mesmo tamanho de frases, conforme apresentado na Figura 8 (Grupo A e Grupo B), percebe-se que a medida em que se diminui o valor do suporte, para um mesmo tamanho de frase, a tendência é que a qualidade das frases melhore.

Por outro lado, analisando-se os grupos de experimentos para um mesmo suporte, com diferentes tamanhos de frases, conforme apresentado na Figura 8 (Grupos 1 e Grupo 2), esperava-se que tamanhos de frases maiores produzissem regras de melhor qualidade, porém, os experimentos realizados não foram capazes de comprovar essa hipótese.

4.2.2. Análise dos Resultados

Após a avaliação das regras geradas, o passo seguinte é a avaliação dos resultados obtidos pela aplicação dessas regras. Com o intuito de se deixar os trabalhos de análise mais concisos, utilizou-se como parâmetro apenas as regras geradas no experimento 6. Porém, esse procedimento pode ser replicado para qualquer um dos demais experimentos realizados. A análise também só avalia as regras consideradas como "Boas", dessa forma, o número inicial de 25 regras geradas cai para 15 regras.

Conforme mencionado anteriormente, essa avaliação faz uma análise qualitativa dos resultados obtidos. A seleção do conjunto amostral a ser analisado é feita de forma a identificar aquelas compras com maior probabilidade de terem sido classificadas de forma errônea. Sendo assim, para cada uma das regras classificadas como "Boa", foram identificados outliers, considerando-se três atributos para o cálculo desses outliers: "Valor unitário", "unidade de medida" e "marca". A escolha desses atributos se deu pelo fato de que produtos iguais tendem a possuir valores unitários próximos, unidades de medida e marca semelhantes, ou seja, caso um determinado produto apresente valor, unidade de medida ou marca muito diferente dos demais produtos do mesmo tipo, ele será classificado como um outlier, e terá uma probabilidade maior de ter sido identificado de forma errônea, sendo assim selecionado para uma análise mais minuciosa. No entanto, cabe ressaltar que tal consideração é apenas uma premissa, visto que, é possível que produtos iguais tenham marcas, unidades de medida ou valores unitários totalmente diferentes. Dessa forma, esse é apenas um critério para a seleção da amostra a ser analisada, porém, a qualidade dos resultados é verificada pela análise individualizada das compras selecionadas.

Quanto a obtenção dos atributos considerados no processo de clusterização, o campo "Valor unitário" já estava presente na base de dados utilizada, como um atributo estruturado. Os campos "unidade de medida" e "marca" foram obtidos durante a etapa de pré-processamento, descrita na Seção 3.1.

O algoritmo utilizado para a detecção dos outliers foi o apresentado em (RAMASWAMY; RASTOGI; SHIM, 2000). No algoritmo em questão, a detecção de outliers é feita através do cálculo da distância de um ponto a seus q-ésimos vizinhos mais próximos. Dessa forma, cada ponto é classificado com base na sua distância a seus q-ésimo vizinhos mais próximos, e os r pontos superiores, neste ranking (ou seja, os r pontos com maiores distâncias a seus vizinhos) são declarados como outliers. Logo, os valores de q e r podem ser definidos como o número de vizinhos a serem considerados e o número de outliers a serem identificados, respectivamente.

Tabela 3 - Outliers identificados por regra

Purchase Identification		Applied rule		Considered Attributes		
CodEmpenho[6]	**Seq**	**Regra**	**Product**	**Value in R$**	**Measurement Unit**	**Brand**
2015NE800889	1	R_2	Diesel	100.000.000,00	Litro	Petrobras
2015NE800586	1	R_2	Diesel	87.279.000,00	Litro	Petrobras
2015NE800803	1	R_3	Água mineral	68.611,32	Garrafao 20 L	Seiva
2015NE801154	1	R_3	Água mineral	43.314,89	Garrafao 20 L	Seiva
2015NE804597	1	R_4	Banana	16.388,00	QUILOGRAMA	CEASA
2015NE803169	1	R_4	Banana	15.000,00	QUILOGRAMA	CEASA
2015NE801459	12	R_6	Produto perecível	9,65	Kg	frigolaste
2015NE801459	8	R_6	Produto perecível	21,46	Kg	sabadini
2015NE801077	1	R_10	Gás liquefeito - glp	63.982,92	KG	GASBALL
2015NE801613	1	R_10	Gás liquefeito - glp	120.000,00	KG	GASBALL
2015NE800466	1	R_11	Água mineral	52.602,06	GALAO 20,00 L	calogi
2015NE800899	1	R_11	Água mineral	61.800,48	GALAO 20,00 L	Hydrate
2015NE800375	1	R_12	Bequer de vidro	350,00	UNIDADE	Leica Biosystems
2015NE806768	1	R_12	Bequer de vidro	900,00	UNIDADE	VELP
2015NE800075	2	R_14	Proveta de vidro	7,00	UNIDADE	rav
2015NE800736	1	R_14	Proveta de vidro	4,60	UNIDADE	Uniglass
2015NE800068	4	R_15	Gasolina Comum	553.834,29	Litros	xxxxxxxxxx
2015NE800809	1	R_15	Gasolina Comum	7.157.000,00	Litro	SHELL
2015NE800588	1	R_16	Balão volumétrico para laboratório	530,00	UNIDADE	-
2015NE802470	1	R_16	Balão volumétrico para laboratório	615,88	UNIDADE	DIOGOLAB
2015NE800001	16	R_19	Resistor filme metálico	0,02	UNIDADE	RohmRohm
2015NE800001	34	R_19	Resistor filme metálico	0,02	UNIDADE	RohmRohm
2015NE800042	7	R_22	Caneta esferográfica	1,10	CAIXA 1.200,00 UN	esferografica
2015NE800006	2	R_22	Caneta esferográfica	135,10	CAIXA 12,00 UN	slider
2015NE800089	2	R_23	Álcool etílico hidratado combustível	78.348,81	LITRO	xxxxxxxxxx
2015NE800549	3	R_23	Álcool etílico hidratado combustível	129.552,00	LITRO	IPIRANGA
2015NE800068	3	R_24	Álcool anidro combustível	553.834,29	Litros	xxxxxxxxxx
2015NE800876	3	R_24	Álcool anidro combustível	148.823,93	Litros	NACIONAL
2015NE800035	1	R_25	Peça para automóvel	560.000.000,00	MENSAL	Conforme Edital
2015NE800134	1	R_25	Peça para automóvel	1.000.000.000,00	UNIDADE - PECAS	ORIGINAL

Fonte: Elaboração do Autor

[6] O código do empenho não está completo (com os 23 dígitos) para não possibilitar a identificação da unidade que executou a compra, visto que, esse trabalho não tem foco na área de auditoria, o único objetivo desse procedimento é fazer a identificação de compras consideradas como outliers e que por essa razão possam ter sido classificadas de forma errônea.

Na execução desse estudo, utilizou-se como parâmetro q (número de pontos vizinhos a serem considerados) o valor 10, e o número de outliers a serem encontrados foi definido como sendo 2 (r =2). Sendo que, para o cálculo da distância entre os pontos foi utilizada a fórmula da distância euclidiana.

Na Tabela 3 são apresentados os resultados obtidos pelo algoritmo para cada uma das regras analisadas. As 2 primeiras colunas trazem a identificação da compra, a terceira e quarta coluna apresentam a regra utilizada e o nome do produto identificado pela regra, respectivamente, e as demais colunas apresentam os atributos utilizados para a detecção dos outliers.

Após a definição da amostra com os registros com maiores probabilidades de terem sido classificados erroneamente em cada uma das regras selecionadas (outliers), o passo seguinte é a análise de cada um desses registros de forma individualizada, nas páginas do Portal da Transparência, a fim de verificar se esses registros realmente correspondem ao mesmo grupo de produtos que a regra se propõem a identificar, ou se eles se referem a outros tipos de produtos e consequentemente tenham sido classificados de forma errada pelas regras.

Sendo assim, todos os 30 registros apresentados na Tabela 3 (2 outliers por regra) foram analisados e verificou-se de forma detalhada os textos descritivos das especificações de compras a fim de se averiguar a que se referia cada uma das compras, possibilitando assim a comparação com o tipo de produto que as regras estavam identificando.

Na Figura 9, são apresentados dois recortes de telas do portal da transparência, cada um com a descrição de uma das compras caracterizada como outliers obtidos na aplicação da regra R_3. Esses recortes de tela permitem verificar que apesar dos valores discrepantes, as especificações das compras realmente se referem ao produto esperado, ou seja, Água Mineral.

Figura 9 - Recortes das telas do Portal das Transparência para registros considerados outliers da regra R_3

Fonte: Elaboração do Autor, com base em recortes de imagens do Portal da Transparência do Governo Federal

4.3. CONCLUSÃO DA AVALIAÇÃO

Nessa análise, não foram identificados produtos classificados de forma errônea, porém, algumas das regras geradas, mesmo sendo consideradas como boas, ainda que classificando as compras de maneira correta, faziam uma classificação muito genérica, que dependendo da finalidade para que se deseje utilizar o processo de mineração de frases desenvolvido, possa não

atender aos objetivos por completo. Exemplos desses tipos ocorrem com as regras R_6 e R_25, que identificam as compras de "produto perecível" e "peça para automóvel", respectivamente, ou seja, uma forma muito genérica para se identificar um produto.

Para evitar situações como estas, a proposta prevê uma etapa opcional, apresentada na Seção 3.6 (refinamento de regras), em que especialistas podem, dentre outras atividades, ajustar as regras de acordo com o grau de especificidade que se deseja dar ao produto.

Os resultados desses experimentos também exemplificam outra atividade que pode ser desempenhada por especialista nessa etapa opcional, que é a junção de regras que identificam o mesmo tipo de produto. Por exemplo, as regras R_3 e R_11 levam ao mesmo produto (água mineral), e por isso poderiam ser agrupadas. Também dependendo do grau de especificidade que se deseja dar ao processo, os especialistas poderiam juntar as regras R_23 e R_24, que identificam respectivamente "álcool etílico hidratado combustível" e "álcool anidro combustível" para levarem a um único produto, um pouco mais genérico, denominado "álcool combustível".

Dessa forma, pode-se concluir que a questão de pesquisa "como identificar de forma automatizada os produtos a partir das especificações textuais que são usadas para caracterizá-los nas descrições dos gastos que são apresentados nos portais de transparência pública?" pode ser respondida pela solução proposta e que a hipótese: "se forem identificadas as sequências de tokens que mais se repetem em um determinado conjunto de descrição de compras, então, essas sequências de tokens caracterizarão os produtos mais comprados desse conjunto de descrições" é verdadeira.

5. APLICAÇÕES

Uma vez identificado a que produto cada uma das descrições de compras se refere, uma série de outras análises tornam-se viáveis. O objetivo dessa seção é apresentar algumas aplicações possíveis com as informações obtidas a partir do emprego das técnicas desenvolvidas nessa pesquisa. Para isso, utilizou-se os dados de itens de empenho, referentes ao período de janeiro de 2016 a junho de 2019, a fim de se fazer a identificação dos produtos que estão sendo especificados de forma textual. Dessa forma, analisou-se 18.492.622 registros de itens de empenho.

5.1. CÁLCULO DE PREÇOS DE REFERÊNCIA DOS PRODUTOS COMPRADOS PELA ADMINISTRAÇÃO PÚBLICA

Conforme sugerido em (CARVALHO et al., 2013), a partir do momento em que se tem os produtos devidamente identificados, torna-se possível se propor preços de referência para os diversos produtos que são comprados pela Administração Pública Federal e estão sendo apresentados em portais de transparência.

Na Tabela 4 são apresentados os preços de referência de 10 produtos, obtidos a partir dos itens de empenhos dos anos de 2016, 2017, 2018 e 2019.

Para o cálculo dos preços de referência apresentados na Tabela 4, considerou-se os preços dos produtos como sendo a mediana dos valores unitários pagos em cada uma das compras desses produtos apresentados no Portal da Transparência, visto que, essa métrica está menos suscetível a influência de outliers.

Outro fator relevante, é que em muitas situações o preço de um produto pode ser influenciado por questões de sazonalidade e de localidade. Porém, como utiliza-se informações

dos empenhos, pode-se considerar qualquer um dos atributos do empenho para se fazer a agregação e definir o critério de formação do preço de referência, como por exemplo, por data, por região, por órgão de governo e etc.

Tabela 4 - Amostra de preços de referência calculados

Produto	Unidade	Preço Referência (Mediana)			
		2016	2017	2018	2019[7]
Água Mineral	Galão 20 L	10,00	8,77	R$ 8,90	8,47
Caneta Esferográfica	Caixa 50 Unidades	36,00	32,22	20,95	25
Caneta Marca Texto	Unidade	0,93	0,88	0,83	0,89
Carne de Boi	Kg	20,80	17,23	17,00	17,25
Carne de Frango	Kg	9,84	7,90	7,92	7,24
Cartucho Tinta Impressora	Unidade	80	79	95,22	81,00
Diesel	Litro	3,73	3,40	3,69	3,95
Gasolina	Litro	4,12	4,10	4,73	4,85
Laranja	Kg	2,48	1,98	2,25	2,41
Papel A4	Resma	14,12	14,62	14,39	14,70

Fonte: Elaboração do Autor, com base em dados tratados do Portal da Transparência do Governo Federal

5.2. IDENTIFICAÇÃO DE COMPRAS COM PREÇOS MUITO ACIMA DO ESPERADO

A partir do momento em que se consegue estabelecer um preço de referência para os diversos produtos comprados e apresentados nos portais de transparência, torna-se possível também se identificar compras que tenham sido feitas com valores muito acima do esperado.

Na Tabela 5 é apresentada uma amostra de 2 valores muito acima do esperado para cada um dos exemplos de preços de referência identificados na Tabela 4. Essa tabela é apenas exemplificativa, visto que, devido ao grande número de compras apresentadas no Portal da Transparência, o número de compras consideradas muito acima do preço de referência também é elevado.

Cabe ressaltar que esses valores elevados, apresentados na Tabela 5, não são suficientes para se dizer que tenha havido irregularidades nos referidos processos de compras, visto que, qualquer indício levantado por meio da análise de dados carece de uma averiguação mais aprofundada por meio de auditorias específicas. Também não faz parte do escopo desse trabalho qualquer tipo de análise de compras individuais. No entanto, todos os processamentos sugeridos nessa monografia tornam viável a identificação de compras que fogem do padrão esperado, possibilitando novos tipos de análises que seriam impossíveis de serem feitas com a informação apresentada em formato original (formato textual).

[7] Os dados do ano de 2019 referem-se ao período compreendido entre os meses de janeiro a junho.

Tabela 5 - Amostra de preços muito acima do esperado de compras realizadas em 2019

Produto	Unidade	Preço de referência (2019)	Número do empenho	Valor Unitário
Água mineral	Galão 20 L	8,47	2019NE800079	70,09
			2019NE800041	320,00
Caneta Esferográfica	Caixa 50 Unidades	25,00	2019NE800044	32,87
			2019NE800029	32,87
Caneta Marca Texto	Unidade	0,89	2019NE800020	12,40
			2019NE800278	15,40
Carne de Boi	Kg	17,25	2019NE800198	53,98
			2019NE800024	48,90
Carne de Frango	Kg	7,24	2019NE800088	32,45
			2019NE800432	28,78
Cartucho Tinta Impressora	Unidade	81,00	2019NE800077	310,80
			2019NE800025	256,20
Diesel	Litro	3,95	2019NE800398	62,50
			2019NE800072	54,94
Gasolina	Litro	4,85	2019NE800473	33,33
			2019NE800211	26,96
Laranja	Kg	2,41	2019NE800151	17,00
			2019NE801032	14,32
Papel A4	Resma	14,70	2019NE801510	160,26
			2019NE800128	142,50

Fonte: Elaboração do Autor com base em dados tratados do Portal da Transparência do Governo Federal

5.3. COMPARAÇÃO ENTRE VALORES PAGOS EM COMPRAS LICITADAS E NÃO LICITADAS

No Brasil há uma lei que prevê que todas as compras realizadas pela Administração Pública devem ser precedidas de licitação, no entanto, essa mesma lei também prever algumas situações em que o procedimento licitatório pode ser dispensável ou inexigível.

Um outro tipo de análise que pode ser feita a partir dos resultados obtidos pela técnica proposta nessa monografia é a comparação entre os preços praticados nas compras de produtos em situações em que houve licitação e nos casos em que esse procedimento não ocorreu.

Na Tabela 6 são apresentados os preços praticados, no ano de 2019, com e sem procedimento licitatório, para os mesmos produtos listados na Tabela 4. Nesse caso, assim como na Tabela 4, os preços foram obtidos pela mediana dos valores unitários de cada uma das compras.

Como pode ser observado na Tabela 5, os produtos tendem a ser comprados por um preço maior quando não há um procedimento licitatório anterior a essa compra. Cabe ressaltar que para o caso das compras de canetas esferográficas e de carne de frango, realizadas no ano de 2019, até o mês de junho, todas as compras foram precedidas do procedimento licitatório.

Tais análises só foram possíveis de serem realizadas pelo fato dos produtos terem sido anteriormente identificados.

Tabela 6 - Amostra de preços praticados em compras com e sem licitação

Produto	Unidade	Preço Praticado (Mediana)	
		Com Licitação	Sem Licitação
Água Mineral	Galão 20 L	R$ 6,20	R$ 10,51
Caneta Esferográfica	Caixa 50 Unidades	R$ 25,00	--
Caneta Marca Texto	Unidade	R$ 0,85	R$ 12,93
Carne de Boi	Kg	R$ 17,03	R$ 20,05
Carne de Frango	Kg	R$ 7,24	--
Cartucho Tinta Impressora	Unidade	R$ 64,45	R$ 105,06
Diesel	Litro	R$ 3,60	R$ 4,03
Gasolina	Litro	R$ 4,36	R$ 4,59
Laranja	Kg	R$ 1,52	R$ 2,99
Papel A4	Resma	R$ 14,07	R$ 17,50

Fonte: Elaboração do Autor, com base em dados tratados do Portal da Transparência do Governo Federal

5.4. IDENTIFICAÇÃO DE FORNECEDORES VENDENDO O MESMO PRODUTO COM PREÇOS DIFERENTES

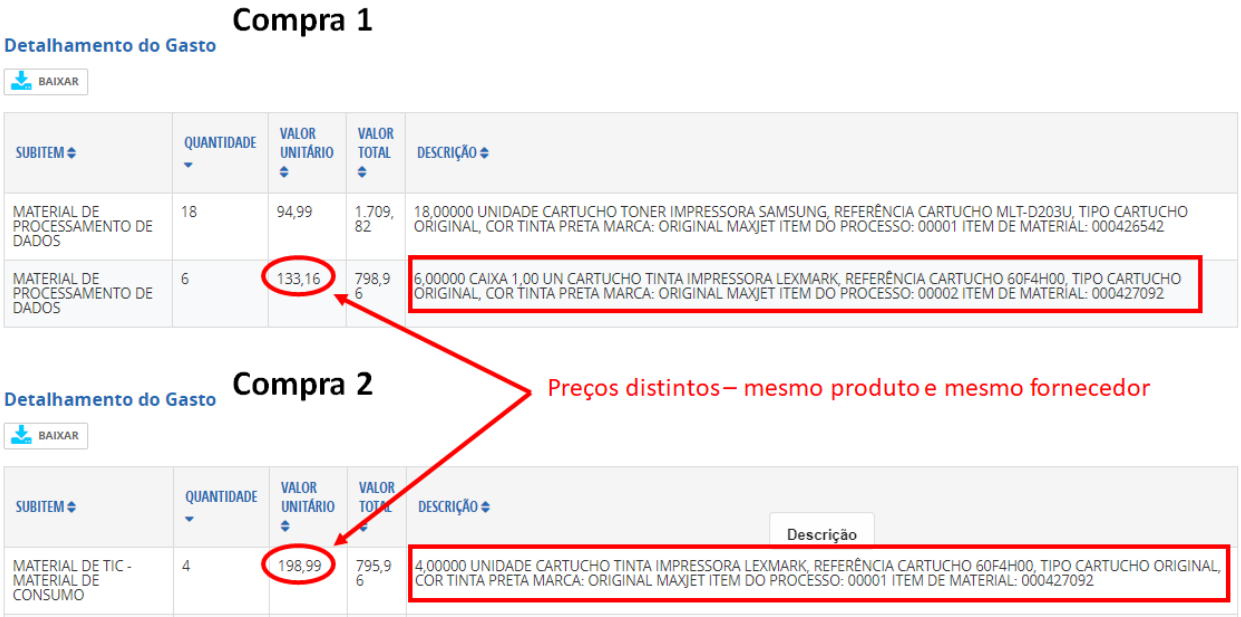
A partir do momento em que se tem os produtos devidamente identificados, pode-se juntar essas informações com outras informações que já estão estruturadas na base de dados do Portal da Transparência, a fim de se enriquecer as análises feitas. Um exemplo disso é a identificação de casos em que o mesmo fornecedor está vendendo o mesmo produto com preços muito diferentes.

Na Figura 9 são apresentados dois recortes de telas do portal da transparência do Governo Federal com duas compras de cartuchos para impressora. Ambas as compras se referem ao mesmo modelo de cartucho e o fornecedor também é o mesmo, porém, o preço a ser pago tem uma variação de mais de 60 %.

Esse caso apresentado é apenas um dos muitos casos semelhantes identificados. Essas disparidades acontecem porque muitas vezes os processos de compra ocorrem de maneira independente, sendo que, uma unidade gestora não fica sabendo do preço que uma outra unidade gestora está pagando pelo mesmo produto ao mesmo fornecedor.

A metodologia ora proposta pode ajudar na otimização das compras realizadas pela Administração Pública, permitindo que a unidade gestora possa propor renegociações de preços a serem pagos, quando identificados valores economicamente mais vantajosos sendo praticados pelo mesmo fornecedor, em outras vendas para a Administração Pública.

Figura 9- Telas do Portal da Transparência com o mesmo fornecedor vendendo o mesmo produto com preços diferentes



Fonte: Elaboração do Autor, com base em recortes de imagens do Portal da Transparência do Governo Federal

5.5. ACOMPANHAMENTO DE TENDÊNCIA DE PREÇOS

Outra possibilidade de aplicação é o acompanhamento de tendência dos preços de um determinado produto e possíveis comparações com os preços praticados pelo mercado nesse mesmo período, para poder se verificar se a Administração Pública está pagando valores condizentes com aqueles praticados pelo mercado.

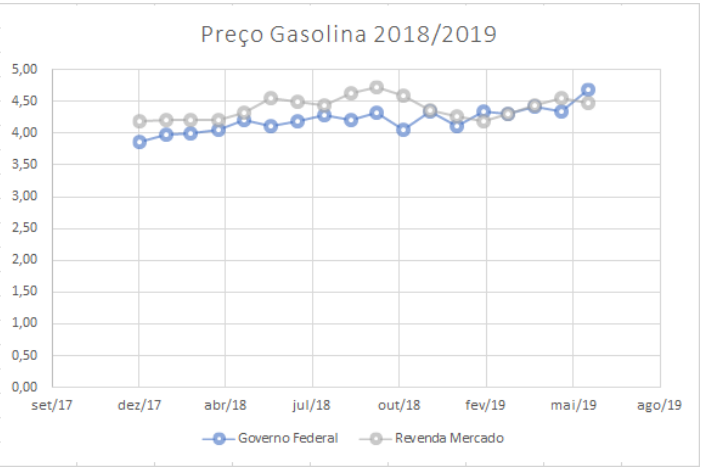
Na Figura 10 é apresentado um gráfico (e a tabela que deu origem a esse gráfico) com um comparativo entre os preços pagos pelo litro da gasolina pelo Governo Federal e o preço praticado pelo mercado[8] no mesmo período (de janeiro de 2018 a junho de 2019).

Pela análise da Figura 10, pode-se concluir que o Governo Federal comprou combustível com preços medianos compatíveis com o valor pago pelo mercado, sendo que, na maioria dos meses ainda pagou um valor ligeiramente inferior. Porém, mais uma vez, só foi possível se fazer esse tipo de análise pelo fato dos produtos terem sido previamente identificados, visto que, a forma como os intes são descritos nos empenhos não permite esse tipo de análise automatizada, e o grande volume de dados impede uma análise manual.

[8] Os valores dos preços praticados pelo mercado para a gasolina foram obtidos no site da Agência Nacional do Petróleo (disponível em http://www.anp.gov.br/precos-e-defesa/234-precos/levantamento-de-precos/868-serie-historica-do-levantamento-de-precos-e-de-margens-de-comercializacao-de-combustiveis).

Figura 10 – Comparativo entre os preços da Gasolina comprada pelo Governo Federal e os valores de mercado

Mês	Governo Federal	Revenda Mercado
jan/18	3,87	4,19
fev/18	3,98	4,21
mar/18	4,00	4,20
abr/18	4,05	4,22
mai/18	4,21	4,31
jun/18	4,12	4,55
jul/18	4,19	4,49
ago/18	4,29	4,45
set/18	4,21	4,63
out/18	4,32	4,72
nov/18	4,05	4,59
dez/18	4,35	4,37
jan/19	4,12	4,27
fev/19	4,35	4,19
mar/19	4,31	4,31
abr/19	4,42	4,44
mai/19	4,35	4,55
jun/19	4,69	4,47



Fonte: Elaboração do Autor, com base em dados tratados do Portal da Transparência do Governo Federal e da Agência Nacional do Petróleo

5.6. OUTRAS APLICAÇÕES

As aplicações apresentadas nessa seção são apenas alguns exemplos de possíveis utilizações para a identificação de produtos a partir da metodologia proposta nesse trabalho. Porém, existe uma série de outras aplicações possíveis, como por exemplo:

- Aplicação de regras de associação para se identificar a probabilidade de um órgão comprar um determinado produto, dado que ele já tenha comprado um conjunto de outros tipos de produtos;
- A identificação dos órgãos que compram com melhores preços e aqueles que pagam mais caro pelos produtos
- A verificação se há algum padrão de comportamento entre empresas fornecedoras de produtos que possa caracterizar algum tipo de conluio ou combinação de preços;
- Identificação das variações de preços praticados nas diferentes regiões do país, e etc.

Os dados gerados durante o processamento sugerido também podem ser integrados com outras bases de dados (governamentais ou não) a fim de se ampliar os tipos de análises a serem feitos.

Logo, as possibilidades de aplicações dos resultados obtidos com os procedimentos propostos nesse artigo são inúmeras, ficando elas limitadas apenas pelas necessidades e criatividade dos analistas de dados que se propuserem a desenvolver estudos com tais informações.

6. CONCLUSÃO

Essa pesquisa propõe um método de descoberta de conhecimento em texto voltado para dados de descrições textuais de compras apresentadas em portais de transparência. Tal método faz

a geração de regras de identificação de produtos por meio da aplicação de um processo de mineração de frases composta de quatro etapas: geração de frases candidatas, filtragem de frases frequentes, poda de sub frases e geração de regras. Sendo que, antes desse processo de mineração de texto propriamente dito, as descrições de compras passam por uma etapa de pré-processamento, que tem o objetivo de preparar o conjunto de dados textuais para o processo de mineração de frases.

O método proposto utiliza um processo de descoberta de conhecimento em texto que recebe como entrada um conjunto de descrições textuais de compras, e oferece como saída um conjunto de regras de identificação de produtos, utilizando três parâmetros de referência: tamanhos mínimos e máximo de frase e suporte mínimo. Opcionalmente, dependendo da disponibilidade de pessoal, pode-se executar uma tarefa adicional, denominada refinamento de regras. Nessa atividade opcional, especialistas podem validar as regras geradas, bem como, adaptá-las de acordo com os propósitos desejados.

A solução proposta foi avaliada utilizando-se os dados de itens de empenho do Portal da Transparência do Governo Federal, referentes ao período de um ano. Sendo que, durante a avaliação verificou-se que a hipótese inicial, que dizia que se forem identificadas as sequências de tokens que mais se repetem em um determinado conjunto de descrição de compras, então, essas sequências de tokens caracterizarão os produtos mais comprados desse conjunto de descrições, era verdadeira.

Logo, a principal contribuição dessa monografia foi a proposta de um método capaz de gerar regras de identificação de produtos a partir de descrições textuais de compras. Porém, outras contribuições intermediárias também resultaram dessa pesquisa, como a proposta de um algoritmo de geração de frases, a proposta de um algoritmo de poda de sub frases e o desenvolvimento de uma metodologia de avaliação dos resultados.

A pesquisa em questão teve como limitação a ausência de um conjunto de dados com as descrições das compras e a respectiva identificação dos produtos a que cada uma dessas descrições se refere (ausência de dados rotulados), o que dificultou a realização de uma avaliação mais objetiva para a verificação dos níveis de precisão atingidos pela técnica desenvolvida. A grande quantidade de registros que compõem a base de dados a ser analisada, torna inviável a identificação de forma manual de um conjunto de dados amostral que seja numericamente significativo para a avaliação dos resultados. Por outro lado, qualquer tentativa de se gerar um conjunto de dados de teste de forma automatizada, o que resolveria o problema do grande volume de dados a ser analisado, poderia utilizar critérios tendenciosos, além do que, esses critérios de identificação também poderiam identificar os produtos de forma errônea, o que distorceria os resultados da avalição realizada.

Por tais razões, optou-se por proceder-se uma validação qualitativa dos resultados encontrados, conforme apresentado na Seção 4. Tal procedimento não foi capaz de informar um valor exato sobre o grau de precisão atingido pela técnica proposta. Porém, foi capaz de indicar que a técnica se mostrou válida para os propósitos para o qual foi concebida. Ainda, de acordo com os critérios estipulados para a seleção das amostras a serem analisadas qualitativamente, não foram encontrados erros que pudesses refutar a validade da técnica proposta.

Como trabalhos futuros, pretende-se aprimorar o método de mineração de texto proposto, e utilizar os resultados obtidos pela aplicação da metodologia desenvolvida, a fim de se realizar novos estudos com os dados originários dos processamentos realizados.

Com relação à possibilidade de melhoramentos do método proposto, pode-se incorporar o procedimento de clusterização, utilizado na Seção 4 para avaliar a qualidade das regras, ao

processo de descoberta de conhecimento em dados textuais proposto, fazendo com que apenas as regras julgadas como boas (ou seja, aquelas cujo processo de clusterização de seus registros não gere um grande número de clusters com registros espalhados por todos eles) sejam consideradas na etapa de geração de regras, melhorando-se assim a qualidade das regras geradas, e consequentemente, aprimorando-se os resultados do procedimento como um todo.

Outro melhoramento que pode ser feito no processo de geração de regras de identificação de itens de compras é a aplicação de análise de similaridades entre os antecedentes das regras a serem geradas, a fim de se eliminar regras cujos antecedentes possuam similaridades superiores a um determinado valor (passado como parâmetro), quando comparada com outros antecedentes de regras geradas.

O procedimento de mineração de frases apresentado nesse artigo também pode ser adaptado para ser utilizado em outros contextos. Por exemplo, esse procedimento pode ser ajustado para identificar as frases mais frequentes em um determinado conjunto de dados textuais, a fim de se levantar quais os tópicos mais relevantes nesse corpus de texto. Tal atividade poderia ser utilizada em análises de textos publicados em redes sociais, blogs, jornais e etc.

Quanto aos estudos que podem ser desenvolvidos a partir dos resultados obtidos pela aplicação das técnicas aqui expostas, existem diversas possibilidades de pesquisas. O processo sugerido consegue obter dados estruturados a partir de um conjunto de dados textuais, ou seja, as descrições textuais de compras são associadas a variáveis categóricas que especificam a que produtos cada uma dessas descrições se refere.

Dessa forma, pode-se utilizar essa nova variável categórica (que identifica os produtos adquiridos) com os demais dados estruturados presentes nos bancos de dados dos portais de transparência, a fim de se aplicar técnicas de mineração de dados para se extrair conhecimentos implícitos a respeito das atividades governamentais que são apresentadas nos portais de transparência pública.

REFERÊNCIAS

BRASIL. (2009). Disponibilização em tempo real de Informações. Recuperado abril 18, 2015, de https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp131.htm

Carvalho, R. N., Sales, L., Da Rocha, H. A., & Mendes, G. L. (2014a). Using Bayesian Networks to Identify and Prevent Split Purchases in Brazil. BMA@UAI 2014 (p. 70–78).

Carvalho, R., Paiva, E. de, Rocha, H. da, & Mendes, G. (2013). Methodology for Creating the Brazilian Government Reference Price Database. X Encontro Nacional de Inteligência Artificial e Computacional, 2013, Fortaleza-CE. Recuperado de http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0033.pdf

Carvalho, R., Paiva, E. de, Rocha, H. da, & Mendes, G. (2014b). Using Clustering and Text Mining to Create a Reference Price Database. Learning and NonLinear Models, 12, 38–52.
CGU, C.-G. (2013). MANUAL da Lei de Acesso à Informação para Estados e Municípios. 1a edição. Brasília: CGU, abr.

CORTES, C.; VAPNIK, V. Support-vector networks. Machine learning, v. 20, n. 3, p. 273–297, 1995.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113.

El-Kishky, A., Song, Y., Wang, C., Voss, C. R., & Han, J. (2014). Scalable topical phrase mining from text corpora. Proceedings of the VLDB Endowment, 8(3), 305–316.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2–3), 131–163.

Hong, H. (2014). The Internet, transparency, and government–public relationships in Seoul, South Korea. Public Relations Review, 40(3), 500–502.

Liu, J., Shang, J., Wang, C., Ren, X., & Han, J. (2015). Mining Quality Phrases from Massive Text Corpora. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (p. 1729–1744). ACM.

Liu, M., Chen, L., Liu, B., & Wang, X. (2015). VRCA: a clustering algorithm for massive amount of texts. Proceedings of the 24th International Conference on Artificial Intelligence (p. 2355–2361). AAAI Press.

Marzagão, T. (2015). Using SVM to pre-classify government purchases. arXiv preprint arXiv:1601.02680.

Paiva, E., & Revoredo, K. (2016). Big Data e Transparência: Utilizando Funções de Mapreduce para incrementar a transparência dos Gastos Públicos. XII Simpósio Brasileiro de Sistemas de Informação, 2016, Florianopolis-SC.

Ren, X., El-Kishky, A., Wang, C., Tao, F., Voss, C. R., & Han, J. (2015). Clustype: Effective entity recognition and typing by relation phrase-based clustering. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (p. 995–1004). ACM.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.

White, T. (2012). Hadoop: The definitive guide. O'Reilly Media, Inc.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: cluster computing with working sets. HotCloud, 10, 10–10.