# IDENTIFICAÇÃO DE PRODUTOS EM DESCRIÇÕES TEXTUAIS DE COMPRAS: uma proposta para portais de transparência pública


## Eduardo Soares de Paiva

Mestre em Informática pela Universidade Federal do Estado do Rio de Janeiro e Auditor Federal de Finanças e Controle da Controladoria Geral da União



**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.*


* Artigo originado do trabalho contemplado com o 1º lugar do Prêmio Ministro Gama Filho - 2019

## 1 INTRODUÇÃO

Visando atender à crescente demanda por informações públicas, o governo brasileiro tem se empenhado em disponibilizar seus dados, tendo inclusive criado legislações específicas, Lei Complementar 131 (BRASIL, 2009), para garantir a disponibilização de dados governamentais. No entanto, 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 foi concebida 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 de a 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 à importância desses gastos) acaba fazendo com que dados relevantes fiquem escondidos no grande volume de informações disponibilizadas, dificultando o entendimento, a comparação e o reúso desses dados.

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

Logo, o objetivo deste artigo 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 usadas para caracterizá-los nas descrições dos gastos apresentados nos portais de transparência pública.

O restante deste trabalho está organizado da seguinte forma: A Seção 2 faz uma revisão de alguns trabalhos relacionados e a Seção 3 apresenta a arquitetura proposta para a solução do problema. Na Seção 4 são apresentadas algumas possíveis aplicações para o método desenvolvido. Finalmente, na Seção 5 é feita a conclusão do artigo.

## 2 TRABALHOS RELACIONADOS

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 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, utilizam-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-chave (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-chave) 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, utilizam-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. O objetivo principal deste trabalho é tentar identificar as compras consideradas suspeitas de terem sido fracionadas, a fim de permitir que providências 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 adquiridos pela Administração Pública. Este 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 desse 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 à 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-chave, 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.

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[1], em um contraponto ao tradicional Bag of Words (SALTON; WONG;

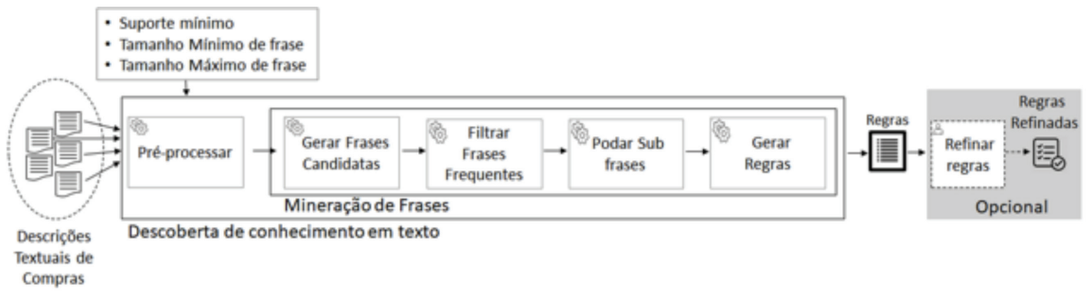
1 Bag of Phrase: modelo de representação utilizado no tratamento de dados textuais. Nesse modelo o texto é representado pela contagem das frases que o compõem, ignorando-se a gramática e a ordem das frases.

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 às 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-chave, por parte de especialistas, para se realizar a extração de informações úteis de conjuntos de dados textuais. Logo, a principal contribuição deste 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.

## 3 PROPOSTA

O método proposto está ilustrado na Figura 1. Ele está dividido em cinco passos obrigatórios e um opcional.

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




Fonte: Elaboração do Autor

No contexto deste trabalho, uma frase é definida como uma sequência contígua de tokens. Sendo assim, neste artigo, 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 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.

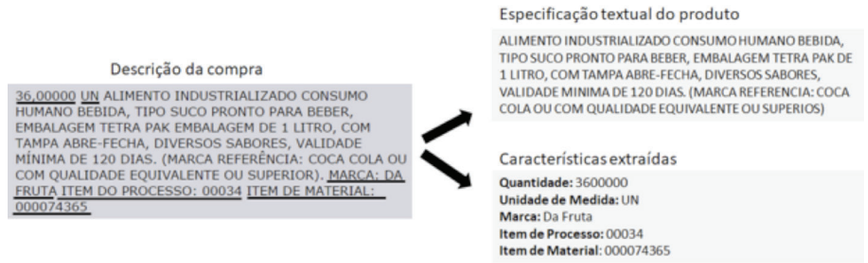
• Completude: Se uma frase longa satisfaz ao critério da frequência, então as subfrases dessa frase longa também irão satisfazer a este critério, porém serão menos informativas do que a frase mais longa, e dessa forma não precisam ser consideradas na mineração, pois a frase mais longa é mais completa.


Devido às 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 com 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**



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, foram 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.

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, pode agregar maior conhecimento a respeito das informações apresentadas.

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 de o 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 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 tokenW está localizado na posição n de uma sequência de tokens de uma sentença, para que esse tokenW 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 Algoritmo1 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

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 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 4é 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 4é 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 superfrases 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 … w_n atende ao requisito do suporte mínimo, então, todas as suas subfrases G^'=w_1 w_2 … 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 o requisito necessário (suporte mínimo). Logo, é executada uma poda aplicando essa propriedade de forma a reduzir o número de frases mineradas.
No Algoritmo 2 é mostrado o processo de poda das subfrases. Esse algoritmo recebe como entrada todas as frases geradas que atenderam ao critério do suporte mínimo, e oferece como saída apenas as superfrases (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 SubFrases**

Algoritmo 2: Poda Sub Frases

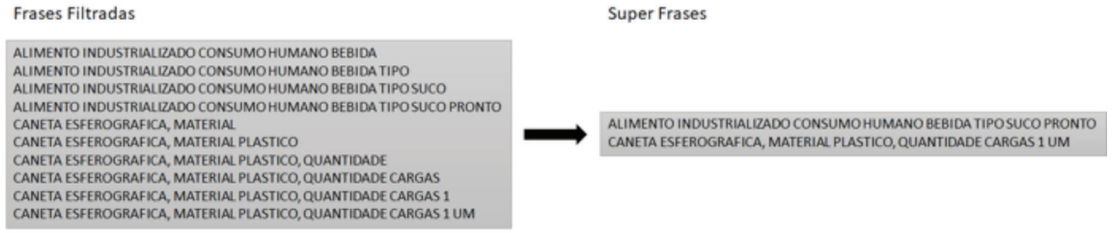
**Entrada:**
Vetor H, com as frases que satisfazem o suporte mínimo
**Saída:**
Vetor com SuperFrases

```
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 subfrases, a Figura 5apresenta, 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 5representa a saída do algoritmo, considerando-se como entrada as frases apresentadas do lado esquerdo da figura.

**Figura 5 - Exemplo do processo de poda de subfrases**

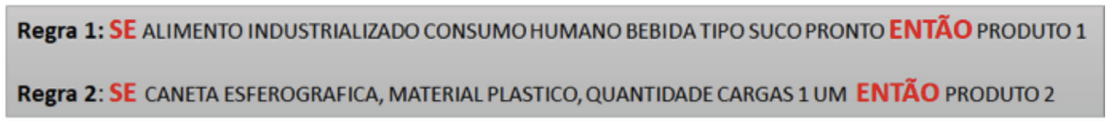



Fonte: Elaboração do Autor

## 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 subfrases 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 superfrases 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. Entretanto, 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 de as frases tenderem a ter um alto grau de expressividade, pois 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 APLICAÇÕES

O objetivo dessa seção é apresentar algumas possíveis aplicações com as informações obtidas a partir do emprego das técnicas desenvolvidas nessa pesquisa. Para isso, foram utilizados os dados de itens de empenho, disponíveis no Portal da Transparência do Governo Federal[2], para período de janeiro de 2015 a junho de 2019. Sendo assim, foram considerados os dados do ano de 2015 para a geração das regras de identificação e os dados dos anos de 2016 a 2019 foram analisados a partir dessas regras. Dessa forma, foram analisados18.492.622 registros de itens de empenho.

## 4.1 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 2016, 2017, 2018 e 2019.
Para o cálculo dos preços de referência apresentados na Tabela 4, foram levados em conta 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 à 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. No entanto, como são utilizadas 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 etc.

2 http://www.transparencia.gov.br
3 Os dados do ano de 2019 referem-se ao período compreendido entre os meses de janeiro a junho.

**Tabela 4 - Amostra de preços de referência calculados**

| Produto | Unidade | Preço Referência (Mediana) | | | |
|---|---|---|---|---|---|
| | | 2016 | 2017 | 2018 | 2019[3] |
| 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, utilizando dados do Portal da Transparência do Governo Federal

## 4.2 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 identificar compras que tenham sido feitas com valores muito acima do esperado.
Na Tabela 5 é apresentada uma amostra com alguns 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 deste trabalho qualquer tipo de análise de compras individuais.

**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 | 2019NE800041 | 320,00 |
| CanetaEsferográfica | Caixa 50 Unidades | 25,00 | 2019NE800044 | 32,87 |
| Caneta Marca Texto | Unidade | 0,89 | 2019NE800020 | 12,40 |
| Carne de Boi | Kg | 17,25 | 2019NE800198 | 53,98 |
| Carne de Frango | Kg | 7,24 | 2019NE800088 | 32,45 |
| CartuchoTintaImpressora | Unidade | 81,00 | 2019NE800077 | 310,80 |
| Diesel | Litro | 3,95 | 2019NE800398 | 62,50 |
| Gasolina | Litro | 4,85 | 2019NE800473 | 33,33 |
| Laranja | Kg | 2,41 | 2019NE800151 | 17,00 |
| Papel A4 | Resma | 14,70 | 2019NE801510 | 160,26 |

Fonte: Elaboração do Autor, utilizando dados do Portal da Transparência do Governo Federal

## 4.3 COMPARAÇÃO ENTRE VALORES PAGOS EM COMPRAS LICITADAS E NÃO LICITADAS

Um outro tipo de análise que pode ser feita a partir dos resultados obtidos pela técnica proposta neste trabalho 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.
Como pode ser observado na Tabela 6, 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.

**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, utilizando dados do Portal da Transparência do Governo Federal

## 4.4 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 já 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 %.
O caso apresentado é apenas um dos muitos casos semelhantes identificados. Essas disparidades acontecem porque muitas vezes os processos de compra ocorrem de maneira independente, porquanto 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**

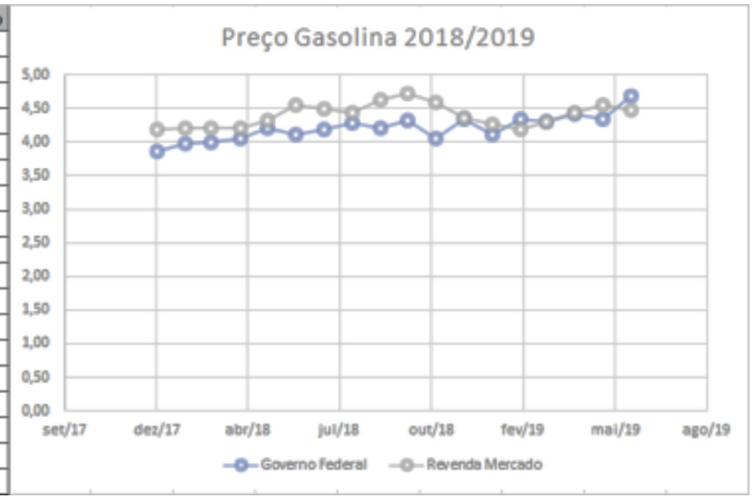

Fonte: Elaboração do Autor, utilizando dados do Portal da Transparência do Governo Federal

## 4.5. ACOMPANHAMENTO DE TENDÊNCIA DE PREÇOS

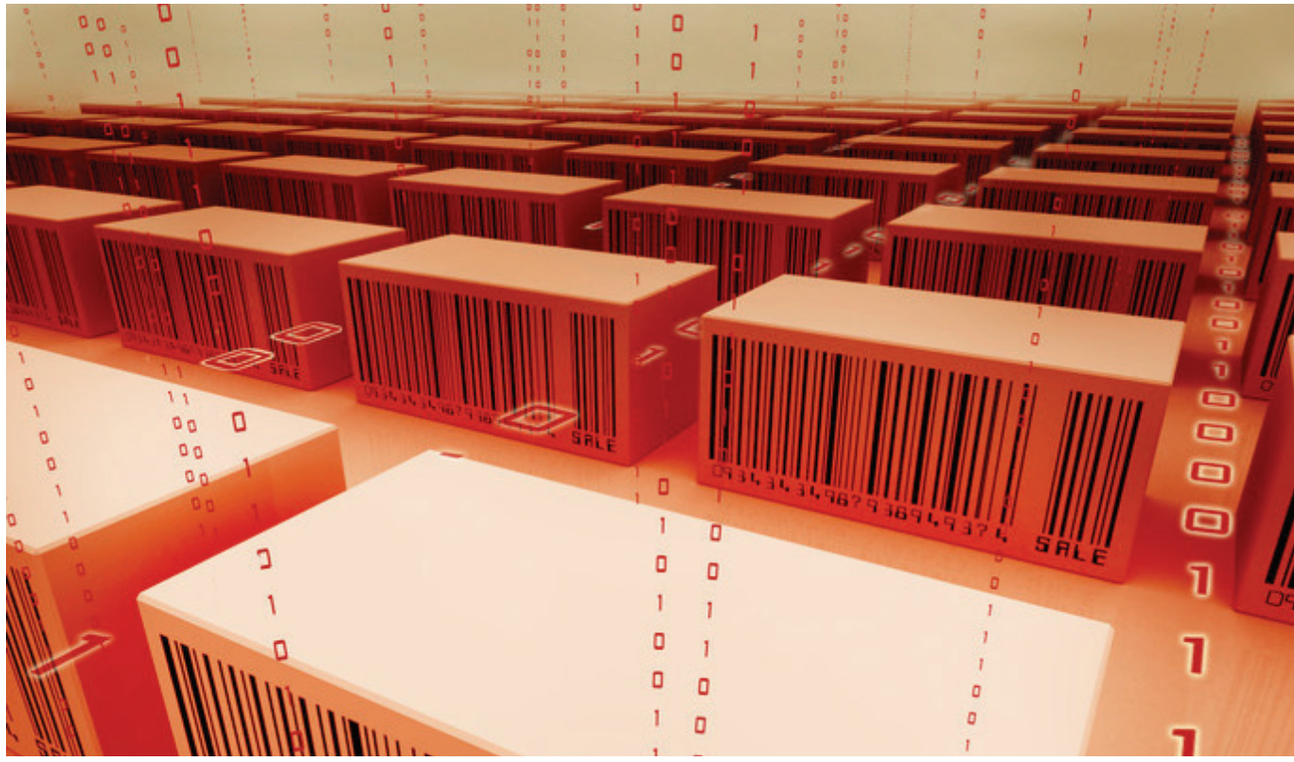
Outra possibilidade de aplicação é o acompanhamento de tendência dos preços de um determinado produto e possíveis comparações com os preços praticados pelo mercado nesse mesmo período, para poder se verificar se a Administração Pública está pagando valores condizentes com aqueles praticados pelo mercado.

**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, utilizando dados do Portal da Transparência do Governo Federal



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[4] 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.

## 4.6 OUTRAS APLICAÇÕES

As aplicações apresentadas nessa seção são apenas alguns exemplos de possíveis utilizações para a identificação de produtos a partir da metodologia proposta neste trabalho. Existe, porém, 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, desde que ele já tenha comprado um conjunto de outros tipos de produtos; identificação dos órgãos que compram com melhores preços e aqueles que pagam mais caro pelos produtos; 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 etc.
Logo, as possibilidades de aplicação dos resultados obtidos com os procedimentos propostos neste artigo são inúmeras, ficando elas limitadas apenas pelas necessidades e criatividade dos analistas de dados que se propuserem a desenvolver estudos com tais informações.

4 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).

## 5 CONCLUSÃO

Esta 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 composto de quatro etapas: geração de frases candidatas, filtragem de frases frequentes, poda de subfrases e geração de regras. 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 mínimo 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.

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.



## 6 REFERÊNCIAS BIBLIOGRÁFICAS


BRASIL. Controladoria-Geral da União. **Manual da Lei de Acesso à Informação para Estados e Municípios**. 1. ed. Brasília, DF: CGU, Secretaria de Prevenção da Corrupção e Informações Estratégicas, 2013.

BRASIL. Lei Complementar nº 131, de 27 de maio de 2009. Acrescenta dispositivos à Lei Complementar no 101, de 4 de maio de 2000, que estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências, a fim de determinar a disponibilização, em tempo real, de informações pormenorizadas sobre a execução orçamentária e financeira da União, dos Estados, do Distrito Federal e dos Municípios. Brasília, DF: Portal da Legislação, [2009]. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/lcp/lcp131.htm. Aceso em: abr18 abr. 2015

CARVALHO, R. N.; SALES, L.; ROCHA, H. A. da; MENDES, G. L. **Using bayesian networks to identify and prevent split purchases in Brazil.** BMAW'14, v. 1218, p. 70–78, 2014a.

CARVALHO, R. N.; SALES, L.; ROCHA, H. A. da; MENDES, G. L. Methodology for creating the brazilian government reference price database. In: Encontro Nacional de Inteligência Artificial e Computacional, 10., 2013, Fortaleza, Ceará. **Anais** [...]. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/eniac/2013/0033.pdf. Acesso em: 18 07 2019

CARVALHO, R.; PAIVA, E. de; ROCHA, H. da; MENDES, G. Using clustering and text mining to carvalho create a reference price database. **Journal of the Brazilian Society on Computational Intelligence (SBIC)**, v. 12, n. 1, p. 38-52, 2014b.

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

EL-KISHKY, A.; SONG, Y.; WANG, C.; VOSS, C. R.; HAN, J. Scalable topical phrase mining from text corpora. **Proceedings of the VLDB Endowment**, v.8, n. 3, p. 305-316, 2014.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine learning**, v. 29, n 2-3, p. 131-163, 1997.

LIU, J.; SHANG, J.; WANG, C.; REN, X.; HAN, J. Mining quality phrases from massive text corpora. **Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data**, p. 1729-1744, may 2015.

LIU, M.; CHEN, L.; LIU, B.; WANG, X. VRCA: a clustering algorithm for massive amount of texts. **Proceedings of the 24th International Conference on Artificial Intelligence**, p. 2355-2361, 2015.

MARZAGÃO, T. (2015). Using SVM to pre-classify government purchases. **arXiv**, Cornell University, 2015. Disponível em: https://arxiv.org/abs/1601.02680. Acesso em: 10 abr. 2020.

PAIVA, E.; REVOREDO, K. Big Data e Transparência: utilizando funções de mapreduce para incrementar a transparência dos gastos públicos. In: Simpósio Brasileiro de Sistemas de Informação, 12., 2016, Florianópolis, Santa Catarina. **Anais** [...]. Florianópolis, SC: Sociedade Brasileira de Computação, maio 2016. Disponível em: https://sol.sbc.org.br/index.php/sbsi/issue/view/364

REN, X.; EL-KISHKY, A.; WANG, C.; TAO, F.; VOSS, C. R.; HAN, J. 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, 2015.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communications of the ACM**, v. 18, n. 11, p.613-620, 1975.

ZAHARIA, M.; CHOWDHURY, M.; FRANKLIN, M. J.; SHENKER, S.; STOICA, I. Spark: cluster computing with working sets. **HotCloud'10: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing**, 10 p., 2010.