Uma ferramenta para recomendação de visualização de dados governamentais abertos

Daiane Macedo[1], **Raissa Barcelos**[1], **Flavia Bernardini**[1], **José Viterbo**[1]

[1]Instituto de Computação Universidade Federal Fluminense (UFF)
Niterói – RJ – Brasil

{daianemacedo, fcbernardini,raissabarcellos, viterbo}@ic.uff.br

Abstract. *The purpose of this work is to present a web tool for recommending visualizations given a set of open data provided. For its construction, a literature review on construction and recommendation of visualizations was carried out. Also, tools that have a similar purpose to the tool developed in this work were analyzed. Based on the literature and related tools, we build models for the data entry and transformation processes, as well as for the decision process of the type of visualization to be suggested. In the conceived scenarios of use, we observe the following advantages of the developed tool: easiness in the creation of visualizations, improvement of the interpretability and contribution for the democratization of the data analysis.*

Resumo. *A proposta deste trabalho é apresentar uma ferramenta web para recomendação de visualizações dado um conjunto de dados abertos. Para sua construção, foi realizada uma revisão da literatura sobre construção e recomendação de visualizações; além da análise de ferramentas que possuem propósito similar à ferramenta aqui desenvolvida. Com base na literatura e nas ferramentas relacionadas, modelamos os processos de entrada e transformação dos dados, assim como o processo de decisão do tipo de visualização a ser sugerido. Nos cenários concebidos de uso, observamos as seguintes vantagens da ferramenta desenvolvida: facilidade na criação de visualizações, melhoria da interpretabilidade e contribuição para a democratização da análise de dados.*

1. Introdução

Com a nova era da informação, o movimento de dados abertos emerge como um fenômeno, com o propósito de tornar os dados, principalmente os dados governamentais públicos, disponíveis para consulta, análise e visualização de dados [Graves and Hendler 2013]. Contudo, o formato no qual os dados ainda são dispostos em portais de dados abertos, em grande volume e em formato tabular, pode parecer confuso para o usuário inexperiente.

O esforço para promover a transparência pública e permitir maior participação do cidadão envolve o uso eficiente de ferramentas on-line para gerenciamento e manipulação de dados. Porém, ainda existe uma lacuna entre a disponibilidade dos dados e o uso efetivo deles, devido à falta de acesso ao hardware ou software necessários, ou mesmo aos recursos e habilidades financeiras e educacionais que permitam o uso efetivo desses dados [Gurstein 2011]. Entre as diversas iniciativas que devem ser tomadas pelos governos para melhorar a interação e a interpretabilidade, que consiste na capacidade de os

usuários entenderem, reconhecerem e interpretarem dados de maneira eficiente e precisa, uma medida possível é o uso de visualizações de dados [Gurstein 2011].

Desde as pinturas nas cavernas, que indicavam onde encontrar alimento, até os gráficos utilizados na revolução industrial, a visualização vem sendo utilizada como um mecanismo para representar dados sobre o mundo, de forma a diminuir a complexidade do processo de transmissão da informação [Berinato 2016]. Logo, a união dos conceitos de dados abertos e visualização de dados promove a inclusão de cidadãos ainda não participantes do processo de compreensão, investigação e sugestões de propostas para os governos baseadas nos dados disponibilizados. Como a construção de uma visualização de dados adequada aos tipos de dados e à intenção do usuário é um processo complexo, se faz necessária uma completa compreensão dos conceitos relacionados à visualização e a dados abertos, que apresentamos na Seção 2. Na Seção 3, apresentamos uma revisão da literatura com recomendações dos autores para o processo de desenvolvimento de visualizações, além da apresentação dos processos realizados por ferramentas de propósito similar ao da ferramenta desenvolvida neste trabalho. Na Seção 4, apresentamos os modelos BPMN elaborados, o mecanismo de recomendação e a ferramenta desenvolvida, que pode ser incorporada nos portais de forma gratuita, tendo em vista que existentes são pagas ou não funcionam como API. Finalizando, a Seção 5 as conclusões deste trabalho e trabalhos futuros.

2. Fundamentação Teórica e Revisão da Literatura

2.1. Visualização de Dados

Linguagens visuais são utilizadas no mundo inteiro e a todo o momento, com o intuito de transmitir informações e auxiliar na compreensão de um cenário ou de tendências futuras. Neste contexto, utilizamos a definição de visualização da informação como a ciência que estuda como exibir dados abstratos visualmente de maneira que tendências, padrões e relações entre estes dados sejam compreendidas e que insights possam ser gerados a partir deles [Nascimento and Ferreira 2005].

Considerando a grande quantidade de métodos de visualização, nos limitamos a abordar os métodos de visualização de dados, que consistem em representar dados quantitativos de maneira estruturada. Partindo desta premissa, gráficos e tabelas são os principais meios de representação de informações da maneira desejada. Cada tipo de gráfico atende a propósitos específicos e é aplicável apenas a determinadas categorias de dados. Concluímos assim que a escolha do melhor gráfico para representar um conjunto de dados, a fim de cumprir os propósitos de uma visualização com qualidade não é uma tarefa simples. Devido à existência de abordagens e técnicas de visualização mais adequadas a um propósito ou a outro, existe a importância de compreender as diferentes visualizações de dados, seus propósitos e aplicações.

Variadas relações entre os dados podem demandar distintas técnicas e abordagens para que os aspectos mais interessantes para o usuário sejam revelados. Dessa forma, os valores inerentes, relacionamentos e estruturas dos dados devem ser considerados no processo de decisão sobre a técnica de visualização mais adequada [Iliinsky and Steele 2011]. Ribecca [Ribecca 2016] cataloga cerca de 60 tipos de visualizações de dados, dentre elas gráficos, diagramas e tabelas, entre outros. Essa variedade dificulta o processo de escolha, assim como estabelece a necessidade de decisão

entre explorar novos tipos de visualizações ou optar pelas visualizações clássicas as quais provavelmente o leitor está mais familiarizado. Já Tufte [Tufte 2001] define os gráficos como sendo frequentemente a melhor maneira de possibilitar a exploração e a explicação dos dados de um dataset. Em complemento aos gráficos, uma propriedade visual utilizada como um recurso efetivo para relacionar valores quantitativos com itens categóricos são as cores, especialmente quando o atributo posição já foi utilizado. Além disso, a utilização de cores funciona também em gráficos de pontos, linhas e barras, desde que o tamanho do objeto não dificulte a distinção das cores. Em gráficos de linhas, quando a distinção categórica por cor já foi utilizada, pode ser realizada a distinção por pontos ou outras formas, que também são eficientes, porém mais difíceis de distinguir que a variação de de cor [Few 2012].

2.2. Dados Abertos e Transparência

Segundo a Open Knowledge Foundation (OKF) [Open Knowledge Foundation (OKF) 20], podemos definir dados abertos como dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa, sujeitos no máximo à exigência de atribuição da fonte e compartilhamento pelas mesmas regras. De acordo com a OKF [Open Knowledge Foundation (OKF) 20], as características que definem os dados abertos são:

- Disponibilidade e acesso: estarem disponíveis completamente e sob custo não maior que um custo razoável de reprodução, de preferência por download pela internet. Os dados também devem estar disponíveis de uma forma conveniente e modificável.
- Reutilização e redistribuição: serem fornecidos sob termos que permitam a reutilização e a redistribuição, incluindo a mistura com outros conjuntos de dados.
- Participação universal: todos devem poder usar, reutilizar e redistribuir. Não deve haver discriminação contra campos de atuação ou contra pessoas ou grupos.

Assim como as organizações privadas, organizações governamentais também produzem uma ampla variedade e grande volume de dados relacionados a diversos âmbitos do governo e da sociedade, como aspectos econômicos e sociais [Ubaldi 2013]. Com a expansão do acesso à informação, as reivindicações pela disponibilização de dados governamentais vêm crescendo progressivamente. Aderindo ao movimento mundial de incentivo à transparência e disponibilização de dados abertos e no intuito de promover uma prestação de contas por parte do governo, o Brasil implantou em novembro de 2011 a lei 12.527, denominada Lei de Acesso à Informação (LAI), que regula o acesso à informação, definindo a transparência, por meio da disponibilização dos dados governamentais, como preceito geral e o sigilo como exceção, exigindo por parte do governo medidas para a disponibilização de dados e suporte aos pedidos de informações. Conceitos como dados abertos, governo aberto, transparência e participação da população são interligados por meio de uma série de recursos e tecnologias, evidenciando a necessidade de ferramentas para auxiliar às pessoas no acesso, compreensão e manipulação desses dados [de Carvalho Freitas et al. 2018].

Em relação aos desafios enfrentados pelo usuário para interagir com os dados já disponibilizados, a utilização tecnologias da informação pode ser uma

solução para tornar o cidadão habilitado a processar, mesclar e interpretar estes dados [Graves and Hendler 2013]. Ferramentas de visualização de dados são exemplos de recursos que possibilitam o consumo e a interação do cidadão com conjuntos de dados, além de facilitar a interpretabilidade e identificação de padrões, integrando assim, cidadãos que não estão habituados a lidar com dados brutos [Barcellos et al. 2017].

2.3. Revisão da Literatura

O desafio de escolher a visualização ideal para representar um determinado conjunto de dados estimulou diversos trabalhos científicos que visam orientar essa escolha, como [Few 2012, Iliinsky and Steele 2011, Berinato 2016, Tufte 2001]. Por outro lado, com o progresso das tecnologias de informação, surgem trabalhos que visam automatizar o processo de escolha e construção da visualização ideal através de ferramentas de recomendação de visualização, como as apresentadas por [Luo et al. 2018, Vartak et al. 2015].

Um agravante do processo de recomendação é o fato de que a classificação de uma visualização como interessante ou não, depende de uma série de fatores, Vartak et al [Vartak et al. 2015] classificam uma visualização como interessante quando ela apresenta uma grande divergência (ou discrepância) a partir de um referencial, que pode ser, por exemplo, outro conjunto de dados, dados históricos ou o restante dos dados presentes no conjuntos de dados fornecido. Ou seja, as visualizações de dados, geradas a partir do conjunto de dados da consulta, que mostram tendências distintas — em comparação com um conjunto de dados de referencial — são consideradas de alta utilidade. Entretanto, Vartak et al [Vartak et al. 2015] apontam que outros fatores podem ser consideradas ao classificar o quão interessante é uma visualização, como a estética, outros atributos particulares dos dados ou algumas tendências existentes.

Sobre o processo de recomendação automatizada de visualizações, Vartak et al [Vartak et al. 2017] definem alguns fatores principais que devem ser considerados ao optar por uma recomendação, destacando que sistemas diferentes podem priorizar diferentes fatores, dependendo do propósito da aplicação. Dentre estes, o aspecto mais relevante a ser considerado é definido pelas características dos dados, uma vez em que o propósito do sistema de recomendação é destacar atributos do conjunto de dados como padrões, tendências e valores de interesse. Entre as características dos dados que um sistema de recomendação deve considerar estão: resumos do conteúdo do dataset, as correlações entre os atributos e os e padrões e tendências existentes.

Outro fator a ser considerado pela ferramenta de visualização é o objetivo da tarefa do usuário como, por exemplo, se tem um caráter mais exploratório ou explicativo, se visa a comparação de subconjuntos de dados a fim de obter insights, ou se busca padrões específicos ou outliers [Vartak et al. 2017]. Mais um ponto a se avaliar é a semântica e conhecimento de domínio, como um fator importante para que as recomendações apresentem informações interessantes para o usuário. A relevância da facilidade de compreensão, buscando recomendar e/ou gerar uma visualização intuitiva e a preferência do usuário também são destacadas por Vartak et al [Vartak et al. 2017] como mais dois aspectos a serem considerados quando busca-se recomendar uma visualização de dados adequada.

2.4. Tecnologias relacionadas

Outras aplicações possuem o propósito automatizar o processo de recomendação de visualização para o usuário. Neste trabalho, avaliamos três ferramentas: SEEDB [Vartak et al. 2015], Tableau Public [Tableau 2003] e DeepEye [Qin et al. 2018].

SEEDB: A proposta do framework SEEDB consiste em, dado um conjunto de dados, verificar as possíveis visualizações que podem ser construídas, avaliar as mais promissoras e recomendar aquelas que podem ser mais úteis para o usuário identificar tendências e outras informações interessantes. Como desafios, os autores citam (a) a escala, devido ao grande número de visualizações que podem ser construídas e (b) a dificuldade de avaliar a utilidade de uma visualização, pois isso depende de uma variedade de fatores. Para a recomendação de visualização, primeiro o usuário especifica na interface os subconjuntos de dados que deseja utilizar. Após, seleciona as colunas do conjunto de dados que devem ser colocadas nos eixos x e y da visualização, como resposta recebe um conjunto de visualizações recomendadas. O diferencial da ferramenta está em exibir um grande conjunto de visualizações ao mesmo tempo em que consegue avaliar se este conjunto contém as visualizações mais interessantes para o usuário.

Tableau Public: O Tableau consiste em um software que permite aos usuários gerar visualizações com dados de diferentes fontes, incluindo arquivos CSV e bancos de dados relacionais. Além disso, também oferece recursos avançados para exploração de dados, como classificação filtragem, drilldown, agrupamento e pivotagem de dados. Por meio da funcionalidade Show Me, o Tableau recomenda o tipo de visualização de dados mais adequado de acordo com os campos de dados de entrada, possibilitando também que o usuário opte por outros tipos de visualizações disponíveis. Para isso, Show Me usa a linguagem de especificação algébrica VizQL. O VizQL executa uma query sobre um conjunto de dados e retorna múltiplas visualizações de dados, como tabelas, gráficos, mapas e séries temporais, dentre outras. O software classifica os campos em dimensões (identificadas na interface pela cor azul), que são tipicamente campos categóricos; e medidas (identificadas na interface pela cor verde), que usualmente são campos quantitativos e passam por algum tipo de agregação. A estrutura básica de uma visualização é definida de acordo com a ordem dos campos inseridos nas áreas “rows” (linhas) e “columns” (colunas) que identificam, respectivamente, as categorias (labels) e o eixo do gráfico recomendado, como mostrado na tabela abaixo [Mackinlay et al. 2007].

As regras de recomendação automática do Tableau baseiam-se nas características dos dados. As três propriedades de dados suportadas pelo software são as seguintes: tipo de dado (texto, data, data e hora, numérico ou booleano); função do dado (medida ou dimensão); e interpretação do dado (discreto ou contínuo). Baseado nessas propriedades dos dados, é realizada a seguinte classificação:

1. C = categórico (dados discretos ou dimensões)
    (a) Cdate = dados categóricos de datas (data ou data e hora)
2. Q = quantitativos (dados contínuos)
    (a) Qd = Quantitativos dependentes (medidas)
    (b) Qi = Quantitativos independentes ou Qdate (dimensões)

Um exemplo para explicar o conceito de dados quantitativos dependente e independentes é: o número que representa o total de vendas de um determinado produto é uma variável

independente, enquanto o valor total recebido pelas vendas é uma variável dependente, pois depende do total de vendas realizadas.

Na Tabela 1, adaptada de Mackinlay et al [Mackinlay et al. 2007], são apresentadas as regras automáticas do modo Automatic Marks, disponível no Tableau, no qual as duas colunas da tabela são referentes aos dados mais à direita nos campos "colunas" e "linhas" presentes da interface do software e, a partir dos tipos de dados, é determinado o tipo de gráfico a ser recomendado.

Campo 1	Campo 2	Tipo de visualização
C	C	tabulação cruzada (crosstab)
Qd	C	Gráfico de barras
Qd	Cdate	Gráfico de linhas
Qd	Qd	Gráfico de dispersão
Qi	C	Diagrama de Gantt
Qi	Qd	Gráfico de linhas
Qi	Qi	Gráfico de dispersão

Tabela 1. Recomendação de visualização do Tableau [Mackinlay et al. 2007]

DeepEye: O software DeepEye, assim com o Tableau e a ferramenta proposta, também leva em consideração em seu processo de recomendação de visualização de dados as propriedades dos dados de entrada. O DeepEye pode ser definido como uma ferramenta de busca que converte as Palavras-chave inseridas pelo usuário no sistema em uma *query*, e, como resultado dessa *query*, apresenta visualizações interessantes para o usuário. O DeepEye mantém um conjunto interno de visualizações classificadas, como boas e segue os seguintes passos para recomendação de visualizações: (i) o usuário insere ou seleciona uma única tabela com dados; (ii) o usuário insere um conjunto de palavras chave para busca nesta tabela; (iii) o mecanismo de busca gera um conjunto de visualizações de acordo com as palavras chave; (iv) o módulo de transformação de visualização realizará a comparação das visualizações geradas com as visualizações armazenadas em um conjunto prévio de boas visualizações e, através dessas comparações, o módulo de ranqueamento de visualizações definirá o ranking com as melhores visualizações; (v) as melhores visualizações, de acordo o software, são retornadas para o usuário; (vi) o usuário pode escolher sua visualização predileta e descobrir mais visualizações pelo módulo de Navegação Facetada.

Para definir qual é a recomendação de visualização para um determinado conjunto de dados, a ferramenta DeepEye considera o tipo de dado de uma coluna, classificado nas seguintes categorias:

- Quantitativos: valores numéricos em uma coluna;
- Discretos: quantidades, sempre representadas por números inteiros;
- Contínuos: valores localizados dentro de uma escala contínua, medidas mensuráveis. Ex: altura, peso, etc.
- Qualitativos: categorias que representam a classificação dos indivíduos. São valores não mensuráveis.

Observamos que a classificação dos dados pelas ferramentas Tableau e DeepEye são equivalentes. A equivalência das classificações dos tipos de dados do Tableau e do

DeepEye são: Categóricos (Tableau) equivalente a Qualitativos (DeepEye); Quantitativos dependentes (Tableau) equivalente a Quantitativos contínuos (DeepEye); e Quantitativos independentes (Tableau) equivalente a Quantitativos discretos (DeepEye).

3. A Ferramenta Proposta

O principal propósito da ferramenta é a recomendação automática de visualizações, permitindo que o usuário selecione quais atributos deseja analisar, priorizando a facilidade de uso. O código está disponibilizado no repositório (Omitido para revisão) e o link para utilização é (Omitido para revisão). As etapas necessárias para a elaboração da ferramenta de recomendação de visualizações de dados são: *definição do escopo da ferramenta*; *classificação dos tipos de dados das colunas do conjunto de dados fornecido*; *transformação dos dados*; e *decisão e construção das visualizações a serem recomendadas para o usuário*.

Para *definição do escopo*, a ferramenta considera que os dados fornecidos estão em formato tabular. Esta delimitação foi estabelecida por tal formato ser um dos mais comuns em conjuntos de dados [Munzner 2014]. No nosso contexto, tabelas são um conjunto de dados formado por itens, que são entidades individuais discretas, como uma linha ou uma célula em uma tabela simples; e atributos, que são propriedades específica que podem ser medidas, como o salário de uma pessoa, por exemplo. Além do formato tabular, a ferramenta aceita como entrada dados em formato CSV por ser um dos formatos mais comuns entre os arquivos disponibilizados nos portais de dados abertos e por atender ao critério de que dados abertos devem ser legíveis por máquinas [Open Knowledge Foundation (OKF) 20 , Umbrich et al. 2015]. Quanto ao gráficos que podem ser recomendados pela ferramenta, fazem parte do escopo os seguintes gráficos estatísticos em 2D: gráficos de colunas, gráficos de linhas, gráficos de pizza e gráficos de dispersão. As características dos formatos citados foram melhor descritas no capítulo anterior sobre gráficos. A delimitação dos tipos de gráficos a serem adotados pela ferramenta se fez necessária, tanto pela necessidade de diminuir a complexidade do processo de recomendação e construção dos gráficos, quanto por alguns formatos não serem recomendados pelos autores consultados [Few 2012, Iliinsky and Steele 2011].

Para *classificação dos tipos de dados fornecidos*, a ferramenta apresentada neste trabalho considera as características dos dados como o aspecto principal da definição do tipo de visualização a ser recomendado. Segundo Munzner [Munzner 2014], muitos aspectos do design de dados são orientados pela natureza dos dados que estão à disposição, englobando tanto a semântica do dado, quanto seus tipos. Partindo desta concepção, a primeira etapa do algoritmo desenvolvido é a análise dos dados fornecidos pelo usuário, classificando cada coluna da tabela individualmente. Os dados das colunas do conjunto de dados fornecido são classificados pela ferramenta em três categorias: dados categóricos, categóricos de data (temporais) ou quantitativos (incluindo quantitativos dependentes e independentes). Essa classificação foi adotada por se mostrar comum nas ferramentas analisadas, porém foi simplificada. Na Figura 1 são exibidos os passos do processo de classificação das colunas.

Para *transformação dos dados*, após a classificação dos dados, aqueles classificados como qualitativos são disponibilizados para seleção no campo “Variável X”, enquanto os quantitativos são exibidos nos campos “Variável X” e “Variável Y”. Essa separação

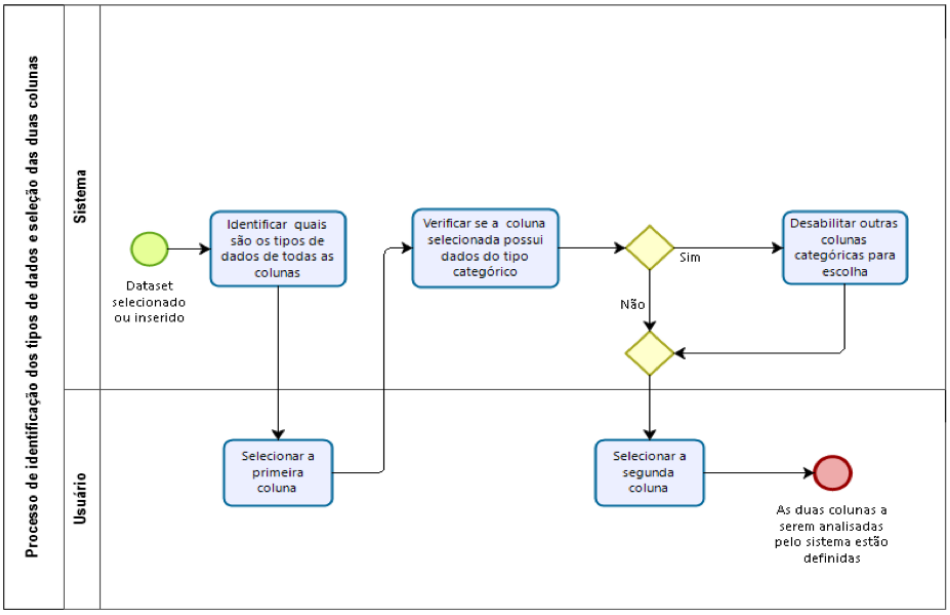

Figura 1. Processo de classificação dos tipos de dados das colunas do conjunto de dados e listagem na interface para seleção, em notação BPMN

ocorre devido à necessidade de existir pelo menos uma coluna de dados quantitativos para a concepção dos gráficos. Além disso, os dados quantitativos são disponibilizados nos dois campos devido à possibilidade de o usuário desejar comparar duas colunas quantitativas, o que ocorreria através de um gráfico de dispersão.

Após a seleção das colunas pelo usuário, que indica o interesse na comparação das colunas na visualização recomendada, ocorre o processo de transformação dos dados por meio das seguintes operações [Qin et al. 2018]:

- Compartimentação (binning): consiste no processo de separar o dado em categorias menores. Valores temporais são particionados em dia, mês e/ou ano. Valores numéricos são particionados com base em intervalos consecutivos, por exemplo: bin1 [0, 10), bin2 [10, 20].
- Agrupamento: consiste em agrupar os valores quantitativos baseados nos valores categóricos correspondentes.

Após a transformação, que categoriza os dados em grupos, os dados de um mesmo grupo podem passar pela etapa de agregação, na qual são executadas funções como a soma, seleção do menor valor, seleção do máximo valor e contagem de valores em uma coluna. Devido às suas características, o dados que representam datas, que podem ser agrupados em dia, mês e ano, recebem um tratamento diferenciado pela ferramenta aqui proposta, assim como no Tableau, que representam esse dado em uma categoria à parte (Cdate), diferenciando-se dos demais dados categóricos. O tratamento consiste no algoritmo identificar quais campos compõem a data, que pode ser formada por dia mês e ano, somente dia e mês ou somente ano; o formato da data, que pode variar entre dia-mês-ano, ano-mês-dia ou mês-dia-ano; e quantos anos distintos estão representados nos dados. Por meio das características dos dados, o nosso processo define se os dados serão agrupados em relação ao ano, em relação ao conjunto mês-ano ou ao mês-ano-dia. Esse processo de compartimentação e agrupamento de dados em relação à data corresponde à operações

de análise denominadas *drill-down*, que consiste em visualizar a informação em nível mais detalhado (menor granularidade) [Van Der Aalst 2013], neste caso, no nível mês e dia; ou rollup, que consiste na agregação do dado em relação a uma ou mais dimensões (maior granularidade) [Van Der Aalst 2013], neste caso, agrupar os dados em relação ao ano correspondente. Os dados de uma coluna de dados quantitativos, além de, caso uma coluna de data tenha sido selecionada, serem agrupados em relação à data, sempre serão agrupados também a em relação aos dados categóricos correspondentes.

Para *recomendação da visualização*, o gráfico a ser construído é determinado pelas colunas selecionadas e pelas características de cada gráfico [Few 2012, Iliinsky and Steele 2011], sendo que esses devem se adequar aos critérios que todos os gráficos devem satisfazer para cumprir o propósito de transmissão da informação — critérios como: (i) exibir os dados; (ii) induzir o espectador a pensar sobre o conteúdo contido no dado, não sobre a metodologia, o design ou a tecnologia para a produção do gráfico; (iii) evitar distorções que o dado não contém; (iv) apresentar muitos números em um pequeno espaço disponível; (v) tornar grandes conjuntos de dados coerentes; (vi) encorajar o leitor a comparar diferentes pedaços do conjunto de dados; (vii) exibir os dados em muitos níveis de detalhamento, de uma visualização ampla até uma granularidade mais fina; (viii) servir a um propósito, descrito claramente; (ix) e ser integrado com a descrição estatística do dataset [Tufte 2001]. No caso do gráfico de pizza, foi encontrada uma contradição, visto que sua utilização é desaconselhada por Few [Few 2012], ao mesmo tempo que Iliinsky e Steele [Iliinsky and Steele 2011] não desaconselham mas recomendam evitar a representação de um gráfico desse tipo com muitas categorias (fatias) para evitar uma visualização confusa. Sendo assim, a solução adotada foi não recomendar o gráfico de pizza quando ele compreender mais de 5 categorias, além disso, o gráfico de pizza é recomendado juntamente com o gráfico de colunas, para que o gráfico de pizza possibilite a comparação do tipo parte-todo, enquanto o gráfico de colunas viabiliza a comparação das categorias de dados entre si. O gráfico de colunas é recomendado pela ferramenta quando o subconjunto de dados consiste em uma coluna categórica e outra quantitativa. O gráfico de linha, como utilizado usualmente, representa a variação de valores (dados quantitativos) ao longo do tempo (coluna de data). Já o diagrama de dispersão, é recomendado pela ferramenta para a comparação de valores quantitativos. Na Figura 2 são apresentados os gráficos a serem recomendados de acordo com os tipos dos dados das duas colunas selecionadas pelo usuário.

4. Conclusões e Trabalhos Futuros

Neste trabalho, apresentamos uma discussão sobre a dificuldade de interpretabilidade de conjuntos de dados abertos, principalmente para usuários sem conhecimentos técnicos sobre o tema. Apresentamos também alguns trabalhos da literatura que propõem o uso de ferramentas de recomendação de visualização como possível solução, bem como uma análise de soluções tecnológicas apresentadas para esse fim. No entanto, são ferramentas que são de difícil incorporação em portais de dados abertos. Assim, propomos uma ferramenta web para recomendação de visualizações. A ferramenta busca implementar as recomendações encontradas na literatura para a construção de visualizações de qualidade, com as características dos dados fornecidos sendo o principal fator direcionador do processo de recomendação. Um dos aspectos positivos da ferramenta consiste em facilitar a criação de visualizações. Um fator relevante é a possível melhora da interpretabilidade,

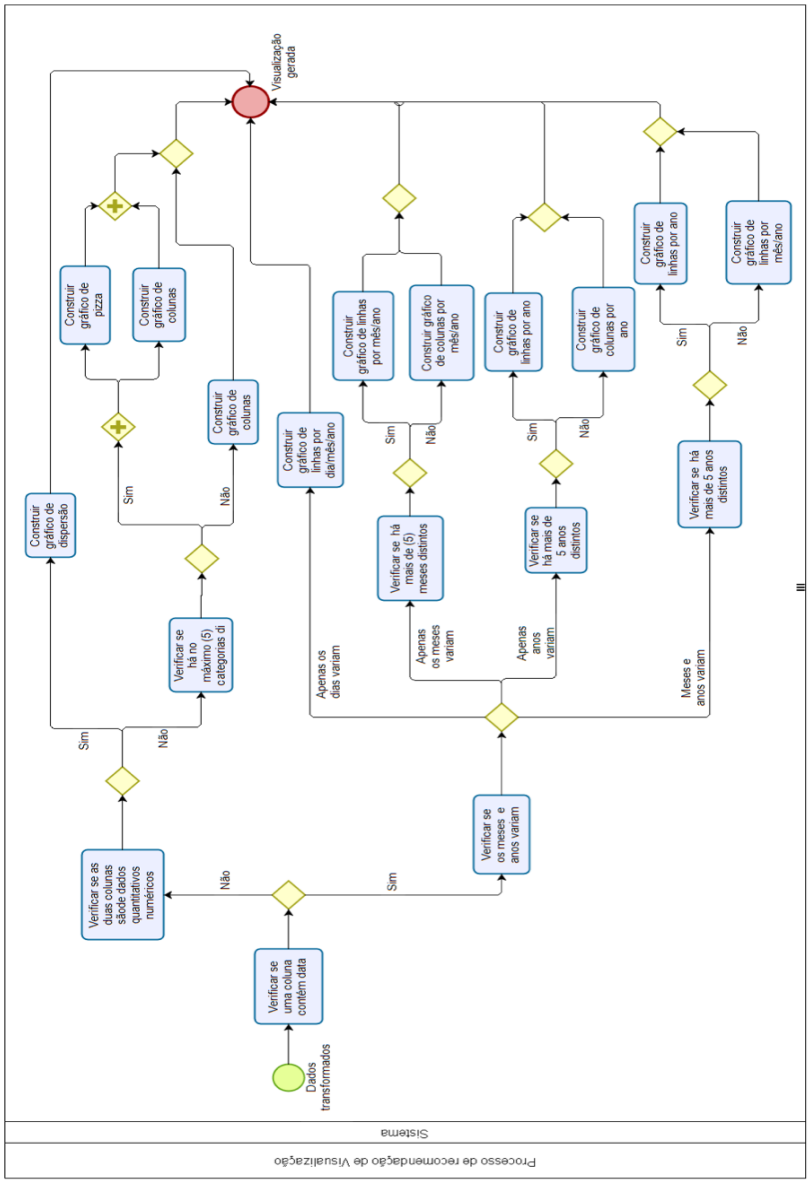

Figura 2. Processo de recomendação de visualização com base nos dados em notação BPMN.

pois sendo baseada na aplicação de técnicas existentes na literatura que visam a escolha das melhores codificações para cada tipo de dado, as visualizações geradas possuem foco na simplicidade, eliminando o problema de visualizações poluídas com características que desviam a atenção do usuário ou dificultam a interpretabilidade. A ferramenta também contribui para a democratização da análise de dados, pois usuários inexperientes nesse assunto podem, com poucos cliques, receber uma recomendação de visualização que evidencia correlações, padrões e tendências, proporcionando insights relevantes. Além disso, a proposta é de uma ferramenta gratuita, enquanto outras opções são pagas. Ao integrar a ferramenta a um portal de dados abertos, por exemplo, um passo crucial é dado em direção à popularização da utilização de dados governamentais abertos.

Quanto a limitações, ainda há pontos que podem ser aprimorados e novas fun-

cionalidades que podem ser implementadas para melhor performance e maior qualidade das visualizações. Primeiramente, é necessária uma análise da arquitetura do sistema e de infraestrutura para a diminuição das taxas de resposta da ferramenta desenvolvida. Uma solução comumente adotada é a utilização de paralelismo para processamento de grandes volumes de dados em menor tempo. Berinato [Berinato 2016] ressalta que o principal critério que define a qualidade de uma visualização é se ela cumpre com o propósito pretendido pelo usuário. Portanto, ainda é necessária uma avaliação com um conjunto de usuários para avaliar a qualidade das visualizações geradas. A amostra selecionada deve representar os perfis de usuários presentes nos ecossistema de dados governamentais abertos, uma referência são os seguintes perfis identificados por Graves e Hendler [Graves and Hendler 2013]: funcionários do governo; consumidores dos dados governamentais (pessoas cujo trabalho implica o uso de dados governamentais, seja para realizar análises ou gerar propostas para futuras políticas públicas); pesquisadores e jornalistas; e cidadãos comuns. Outro objetivo importante é identificar a tarefa ou visão pretendida pelo usuário. Uma possível solução, de baixo nível de complexidade, a ser adotada é a obtenção de informações explicitas sobre a tarefa pretendida através da definição de um conjunto de possíveis atividades como: comparação, correlação, identificação de valores outliers, etc. e a possibilidade de que o usuário escolha uma dessas opções em uma lista na interface da ferramenta. Uma opção mais sofisticada seria o uso de mecanismos de aprendizado de máquina para inferir a intenção do usuário através do comportamento dele ao utilizar o software [Vartak et al. 2017]. Além dos aspectos citados, é ampla a quantidade de estudos publicados sobre sistemas de recomendação e desenvolvimento de visualizações de dados, portanto, a tarefa de encontrar soluções para as limitações desta ferramenta e coleta de possíveis novas funcionalidades não deve apresentar grandes obstáculos.

Referências

Barcellos, R., Viterbo, J., Miranda, L., Bernardini, F., Maciel, C., and Trevisan, D. (2017). Transparency in practice: Using visualization to enhance the interpretability of open data. In *Proc. 18th Annual Int. Conf. Digital Government Research — DGo'2017*, pages 139–148.

Berinato, S. (2016). *Good charts: The HBR guide to making smarter, more persuasive data visualizations*. Harvard Business Review Press.

de Carvalho Freitas, J. A., Balaniuk, R., da Silva, A. P. B., and da Silveira, V. S. (2018). O ecossistema de dados abertos do governo federal: composição e desafios. *Ciência da Informação*, 47(2).

Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. Analytics Press, 2nd edition.

Graves, A. and Hendler, J. (2013). Visualization tools for open government data. In *Proc. 14th Annual Int. Conf. Digital Government Research — DGo'2013*, pages 136–145.

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 16(2).

Iliinsky, N. and Steele, J. (2011). *Designing data visualizations. Intentional communication from data to display*. O'Reilly Media.

Luo, Y., Qin, X., Tang, N., and Li, G. (2018). Deepeye: Towards automatic data visualization. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 101–112.

Mackinlay, J. D., Hanrahan, P., and Stolte, C. (2007). Show me: Automatic presentation for visual analysis. In *IEEE transactions on visualization and computer graphics*, pages 1137–1144.

Munzner, T. (2014). *Visualization Analysis and Design*. CRC Press.

Nascimento, H. A. and Ferreira, C. B. (2005). Visualização de informações–uma abordagem prática. In *XXIV JAI do XXV Congresso da Sociedade Brasileira de Computação*.

Open Knowledge Foundation (OKF) (20–). Open data handbook. Disponível em `https://opendatahandbook.org/guide/pt_BR/what-is-open-data`. Acessado em 28/04/2020.

Qin, X., Luo, Y., Tang, N., and Li, G. (2018). Deepeye: An automatic big data visualization framework. *Big data mining and analytics*, 1(1):75–82.

Ribecca, S. (2016). The data visualisation catalogue. Disponível em `http://www.datavizcatalogue.com/index.html`. Acessado em 28/04/2020.

Tableau (2003). Tableau software. Disponível em `https://www.tableau.com/pt-br`. Acessado em 28/04/2020.

Tufte, E. R. (2001). *The visual display of quantitative information*. Graphics Press.

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance 22, OECD Publishing.

Umbrich, J., Neumaier, S., and Polleres, A. (2015). Quality assessment and evolution of open data portals. In *3rd International Conference on Future Internet of Things and Cloud. IEEE*.

Van Der Aalst, W. M. (2013). Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining. In *Asia-Pacific conference on business process management*.

Vartak, M., Huang, S., Siddiqui, T., Madden, S. R., and Parameswaran, A. (2017). Towards visualization recommendation systems. *ACM SIGMOD Record*, 45(4):34–39.

Vartak, M., Rahman, S., Madden, S., Parameswaran, A., and Polyzotis, N. (2015). SeeDB: efficient data-driven visualization recommendations to support visual analytics. In *Proc. VLDB Endowment*.