

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
MESTRADO 2015.2
VISUALIZAÇÃO DE INFORMAÇÃO
Prof Asla Medeiros e Sá

Projeto e Implementação da Visualização *Explorando o Media Cloud Brasil*

KIZZY TERRA

RIO DE JANEIRO
SETEMBRO DE 2015

1 Introdução

O autor Andy Kirk apresenta em seu livro intitulado *Data Visualization: A successful design process* [1], uma metodologia para a construção de visualizações de dados que compreende uma sequência de tarefas e decisões analíticas as quais permitem que se possa desenvolver visualizações de maneira mais organizada, eficiente e efetiva garantindo uma melhor comunicação da informação a ser visualizada. Nas seções a seguir deste documento são apresentadas as principais escolhas e considerações feitas durante o planejamento da visualização aqui proposta, segundo a metodologia sugerida por Andy Kirk, bem como sua posterior implementação.

2 Planejamento da Visualização

Segundo o autor Andy Kirk, sua metodologia para criação de uma visualização não se trata de um conjunto de instruções que devem ser seguidas de forma dogmática, mas sim um conjunto de diretrizes que devem ser utilizadas de forma flexível para guiar o processo de construção de uma visualização. O autor enumera, basicamente, cinco grandes aspectos [2] que devem ser cuidadosamente pensados e analisados, são eles:

1. Estabelecer o objetivo da visualização e identificar principais parâmetros
2. Adquirir, preparar e explorar os dados
3. Estabelecer o foco da apresentação dos dados (procurar histórias)
4. Conceber o design da visualização
5. Construir a solução de visualização

Nas subseções a seguir, apresenta-se uma proposta de visualização de dados a qual foi idealizada a partir da aplicação do processo de criação sugerido pelo autor supracitado, levando-se em consideração os aspectos mencionados anteriormente.

2.1 Estabelecendo o objetivo

Pretende-se criar uma visualização de dados exploratória a fim de permitir uma análise visual dos dados. Em outras palavras, a visualização deverá ser uma ferramenta a qual irá funcionar como uma interface de exploração de um conjunto de dados (ver seção 2.2).

2.2 Descrição dos dados

Para implementar a visualização aqui proposta serão utilizados os dados capturados pelo Projeto Media Cloud Brasil desde 05/2013 até o presente momento. A base de dados do Media Cloud está hospedada em um banco de dados não relacional, mais especificamente, os dados estão em um servidor MongoDB. Cada artigo capturado e armazenado possui os atributos apresentados na imagem a seguir:

```

{
  "_id" : ObjectId("54d78a65dcccdd4c198c5ca8"),
  "compressed" : true,
  "language" : {
    "code" : "pt",
    "name" : "PORTUGUESE"
  },
  "link" : "http://infograficos.oglobo.globo.com/economia/defesa-do-consumidor/ranki",
  "link_content" : { "$binary" : "eJzVPNTuJDd2m7y1gP0HusaxJFjVpW5JM6Nb0xppZnZ25waNb0",
  "published" : null,
  "source" : "crawler_oglobo",
  "status" : 0,
  "title" : "O Globo - Ranking da Secretaria Nacional do Consumidor"
}

```

Figura 1: Lista de atributos de um artigo do Media Cloud

Inicialmente, serão considerados os atributos *source* e *title* para construir a visualização, posteriormente pode-se considerar também o conteúdo dos artigos. A consulta aos dados que irão compor a visualização será realizada em tempo real, visto que pretende-se que a visualização seja interativa.

2.3 A anatomia da visualização

2.3.1 Esboço da visualização

Primeiramente, apresenta-se um esboço da visualização que se pretende construir, a fim de facilitar o entendimento das seções a seguir as quais discutem as escolhas feitas para a forma de representação e apresentação dos dados.

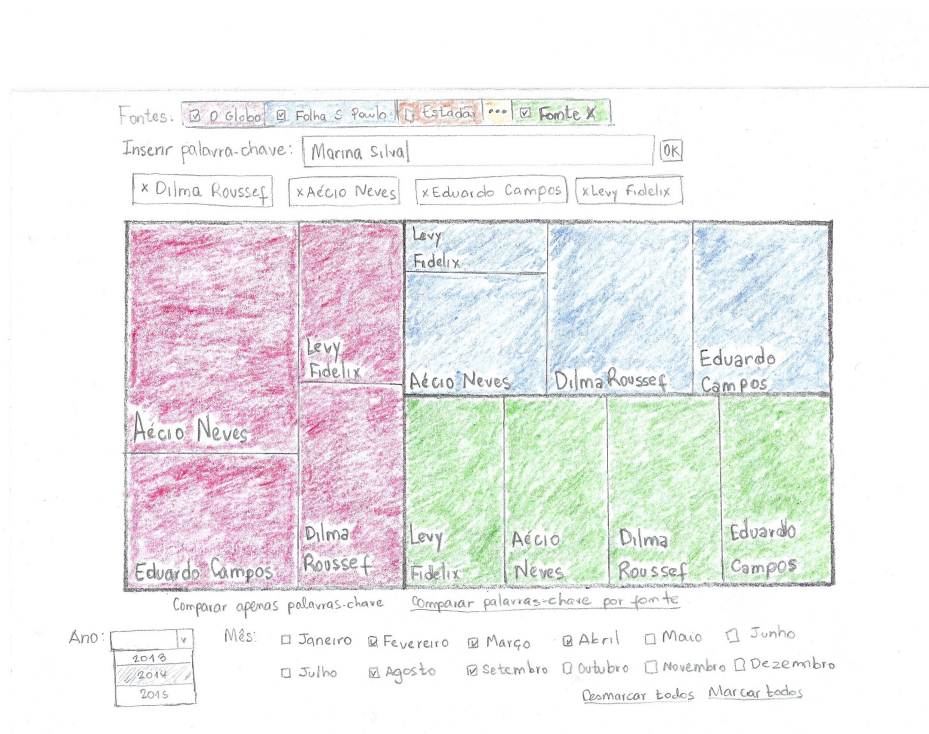


Figura 2: Esboço da visualização proposta - Comparação palavras-chave por fonte

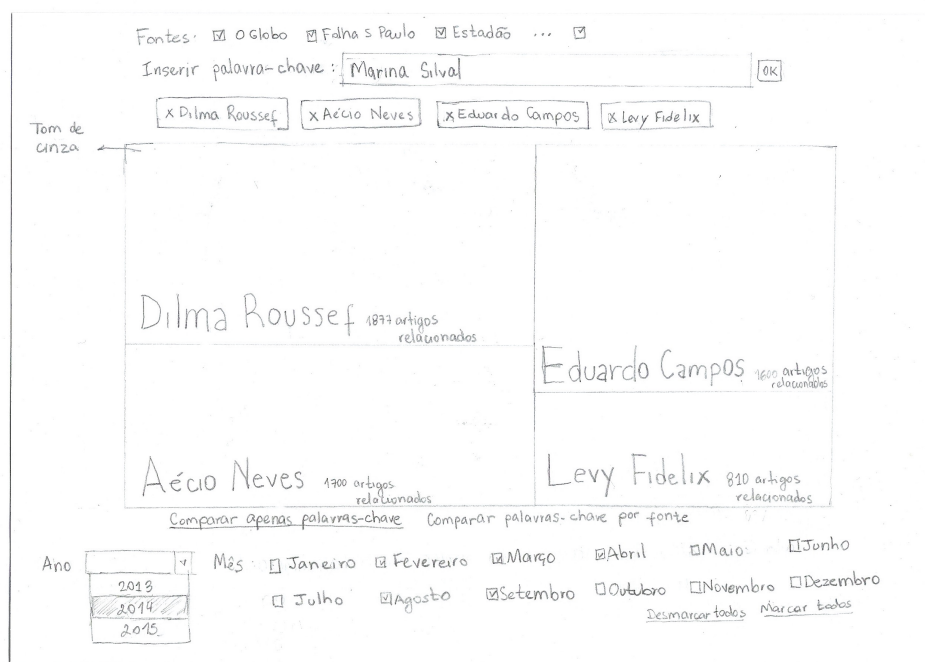


Figura 3: Esboço da visualização - Comparação por palavras-chave

A visualização a ser implementada de acordo com o esboço apresentado nas figuras anteriores possuirá funcionalidades interativas, para bem utilizar a ferramenta o usuário deverá inserir um conjunto de palavras-chave que pertençam a um mesmo contexto e para as quais faça sentido estabelecer-se comparação. Nas figuras apresentadas, por exemplo, mostra-se um exemplo de uso da ferramenta em que todas palavras-chave são políticos que concorreram às eleições de 2014 (*Dilma Rousseff*, *Aécio Neves*, *Marina Silva*, *Eduardo Campos*, *Levy Fidelix*), evidentemente estas palavras referem-se a um mesmo contexto - político - e podem ser tratadas como variáveis categóricas-nominais.

Por outro lado, se ao invés de inserir estas palavras-chave o suposto usuário tivesse inserido palavras-chave desconexas como por exemplo: *Dilma Rousseff*, *Cachorro*, *Bola*, *Laranjas*. A visualização produzida pela ferramenta não iria ter sentido devido ao fato de que estas variáveis não pertencem a um mesmo contexto evidente e portanto, não podem ser tratadas como variáveis categóricas-nominais.

2.3.2 Representação dos dados

A escolha de um design apropriado para os dados, que satisfaça as restrições impostas pela forma como os dados se apresentam, é crucial. Os dados a serem representados na visualização proposta são do tipo categóricos-nominais e quantitativos.

A principal variável quantitativa presente na visualização indica o número de artigos relacionados a uma determinada palavra-chave.

As variáveis categóricas-nominais são as diferentes palavras-chave que podem ser escolhidas pelo usuário e as diferentes fontes de artigos disponíveis para seleção na ferramenta de visualização a ser construída.

2.3.3 Apresentação pretendida para os dados

Segundo a metodologia de Andy Kirk, na etapa de decisão a respeito da apresentação dos dados deve-se fazer escolhas sobre o uso de cor, possíveis funcionalidades interativas, uso de legendas explicativas e o arranjo e arquitetura visual da visualização.

No contexto desta proposta de visualização, duas visualizações já existentes foram tomadas como inspiração: a *Newsmap* [3] e a *Pantheon* [4].



Figura 4: Newsmap

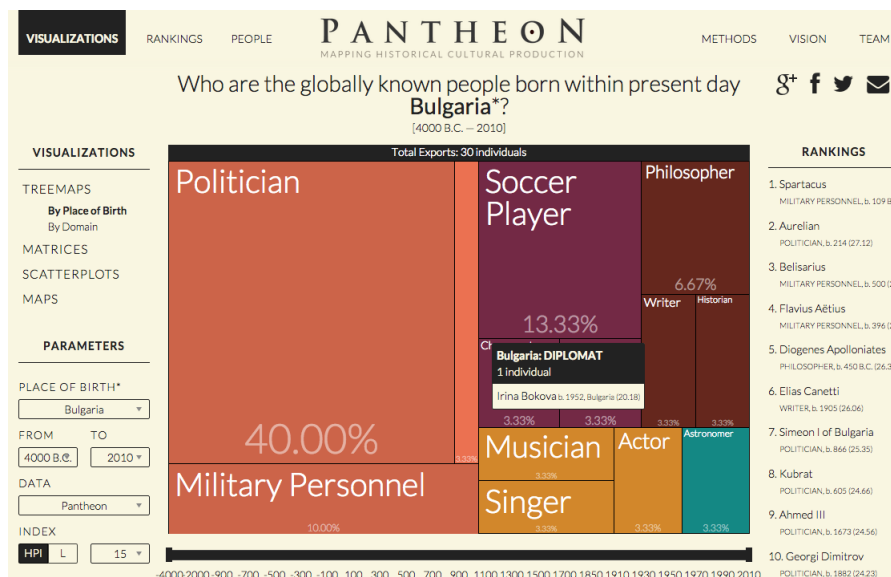


Figura 5: Pantheon- Mapping Historical Cultural Production

Ambas visualizações utilizam um *Tree Map* para apresentar os dados, assim como a visualização que pretende-se criar conforme o esboço apresentado (figuras 2 e 3).

Os tipos de variáveis de dados usualmente representados com *tree maps* são múltiplas variáveis categóricas-nominais e uma variável quantitativa (razão). Além disso, as variáveis visuais exploradas por este tipo de *design* de visualização são: área, posição, intensidade de cor, saturação e brilho

de cor. Dadas essas características, este *design* parece ser bastante adequado para a visualização pretendida.

Assim como as visualizações mencionadas, a visualização proposta também irá possuir funcionalidades interativas: um campo de inserção de palavras-chave e filtros para escolha de fontes, meses e ano. Estas funcionalidades interativas justificam-se por permitirem uma experiência exploratória mais diversificada e dinâmica.

2.3.4 Ferramentas e recursos a serem utilizados

As principais ferramentas a serem utilizadas são: D3.js [5], Python [6], DjanMongoDB [7], PyMongo [8], Bootstrap [9].

3 Implementação da Visualização

Nas seções a seguir são descritas as principais etapas que compreenderam a implementação da visualização planejada anteriormente.

3.1 Metodologia

A metodologia utilizada para implementação da visualização foi *front-to-back*, isto é iniciou-se pelo *front-end* (interface web e *treemap*) da aplicação e após, implementou-se o *back-end* (servidor Django e comunicação com o banco de dados MongoDB).

3.2 Implementação do *front-end*

A interface gráfica da aplicação web foi implementada utilizando-se Bootstrap, HTML, CSS. Devido as dimensões da tela e visando uma apresentação limpa e harmônica dos elementos gráficos na tela o design esboçado apresentado na seção 2.3 foi aprimorado durante a implementação da versão final.

As principais mudanças foram os elementos de seleção de mês e ano que no esboço estavam abaixo do *treemap* e na versão final (figura XXX) estão na lateral esquerda, bem como os elementos de seleção de fonte e as palavras-chave inseridas pelo usuário, que passaram para a lateral direita. Essas modificações possibilitaram que o *treemap* ficasse maior e mais centralizado em relação à página como um todo.

3.2.1 Implementação do *Treemap*

Foram implementados dois *treemaps* utilizando D3.js: um para comparar as palavras-chave por fonte e outro para comparar as palavras-chave apenas . A diferença crucial entre os dois *treemaps* está na cor, visto que no *treemap* que compara apenas palavras-chave o uso de cor deixa de ser necessário, pois não há a diferenciação por fontes. Em função disso, a visualização para a visualização por palavras-chave apenas escolheu uma mesma cor neutra para todas as palavras e utilizou-se apenas a área como representação de informação.

No *treemap* de comparação por fontes e palavras-chave as cores codificam a informação de diferenciação entre as cinco diferentes fontes. A escolha das cores teve como inspiração outras visualizações de *treemap* já implementadas. As cores escolhidas possuem tons mais escuros e são bastante harmônicas entre si.

3.2.2 Dificuldades Encontradas

A principal dificuldade encontrada nesta etapa da implementação foi a necessidade de aprender todas as ferramentas necessárias para então poder iniciar a implementação da visualização.

3.3 Implementação do back-end

3.3.1 Servidor Django

Implementou-se um servidor em Django para comunicar-se com a página web e enviar e receber as informações necessárias para o correto desempenho da ferramenta.

3.3.2 Comunicação com o banco de dados

A comunicação entre o servidor e o banco de dados foi realizada utilizando-se o pacote python PyMongo. Para cada quadrado a ser representado no *treemap*, referente a uma palavra-chave e uma fonte específica, foi implementada uma consulta ao banco, filtrando os resultados pelo período - mês e ano - selecionado pelo usuário.

3.3.3 Implementação da comunicação entre a interface e o servidor

A comunicação entre os elementos da interface e o servidor foi feita utilizando javascript, jQuery e Ajax. Essas ferramentas foram utilizadas para capturar as opções selecionadas pelo usuário (período, fontes e palavras-chave) e enviar para o servidor Django através de um HTTP POST.

3.4 Versão final da visualização

A versão final da visualização é apresentada através das imagens a seguir.

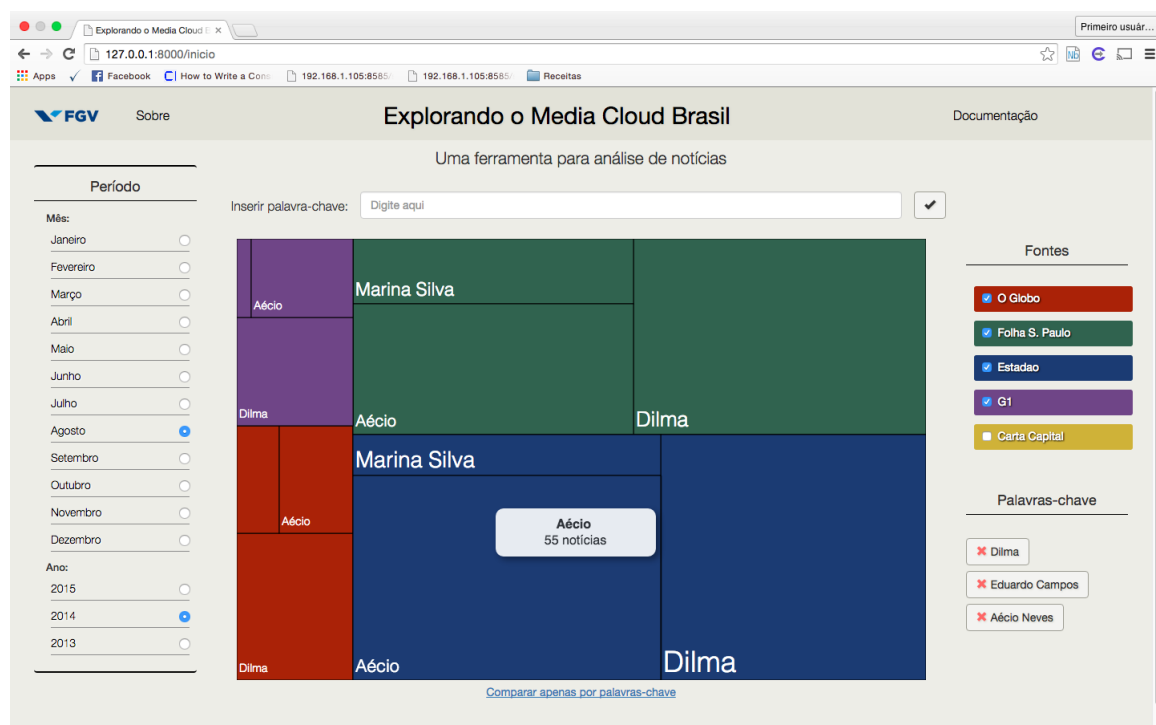


Figura 6: Treemap de comparação de palavras-chave por fonte

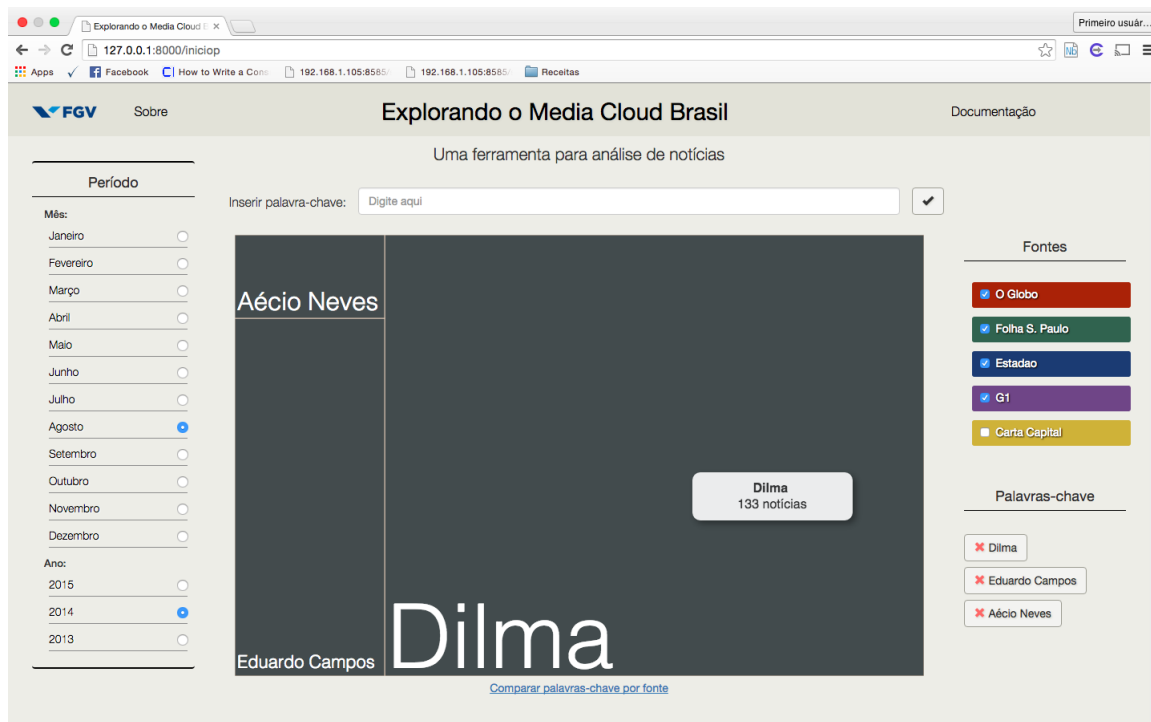


Figura 7: Treemap de comparação de palavras-chave apenas

4 Conclusão

O processo de planejamento e implementação da visualização foi bastante produtivo e proveitoso. Esta foi uma grande oportunidade de aprender muito sobre o que se deve levar em conta para representar informações, bem como foi uma grande oportunidade de adquirir conhecimentos e colocá-los em prática. O resultado atingido foi bastante satisfatório e sem dúvida, foi apenas o início de um trabalho que será ainda mais aperfeiçoado.

Referências

- [1] KIRK, A. *Data Visualization : a successful design process*. [S.l.]: Packt Publishing, 2012. 181 p. ISBN 9781849693462.
- [2] KIRK, A. Visualization Workflow: Finding Stories and Telling Stories. Conjunto de Slides. 2013.
- [3] NEWSMAP. Disponível em: <<http://newsmap.jp>>.
- [4] PANTHEON - Mapping Historical Cultural Production. Disponível em: <<http://pantheon.media.mit.edu>>.
- [5] D3.JS - Página oficial. Disponível em: <<http://d3js.org/>>.
- [6] PYTHON - Página oficial. Disponível em: <<https://www.python.org/>>.
- [7] MONGODB - Página oficial. Disponível em: <<https://www.mongodb.org/>>.
- [8] PYMONGO - Página oficial. Disponível em: <<https://api.mongodb.org/python/current/>>.
- [9] BOOTSTRAP - Página oficial. Disponível em: <<http://getbootstrap.com/>>.