Francisco Morgani Fatore

16 de agosto de 2012

Resumo

Capítulo 1

Introdução

Capítulo 2

Trabalhos Relacionados

Capítulo 3

Metodologia

3.1 Materiais

Java, Postgresql, LibreOffice, InkScape, GIMP, DIA, SchemaSpy, repositório TSE, DATASUS.

Distância de Levinshtein.

3.2 Levantamento dos Dados

Os dados utilizados neste trabalho são provenientes do repositório de dados [1] do TSE (Tribunal Superior Eleitoral). O TSE é o órgão máximo da Justiça Eleitoral brasileira e atua em conjunto com os TREs (Tribunais Regionais Eleitorais), que são os responsáveis diretos pela administração do processo eleitoral dos respectivos estados e municípios os quais representam. [2]

O repositório de dados eleitorais é uma compilação de dados brutos das eleições desde 1994 voltada a pesquisadores, imprensa e cidadãos em geral que tenham interesse em analisar os dados de eleitorado, candidaturas, resultados e prestação de contas das eleições.

Os dados são fornecidos como arquivos de texto formatados no padrão CSV [3] (Comma Separated Values), que é um tipo de formatação muito utilizado devido à sua simplicidade. O

conteúdo de cada arquivo é descrito por arquivos de ajuda. Nestes arquivos de ajuda é comum o uso do termo "unidade eleitoral", a definição deste termo no caso de eleição majoritária é o estado em que o candidato concorre e em caso de eleição municipal é o código TSE do município. Assume-se os valores especiais BR, ZZ e VT para designar as unidades eleitorais correspondentes, respectivamente, ao Brasil, ao Exterior e a Voto em Trânsito.

O TSE não se responsabiliza por qualquer consulta ou análise realizada sobre os dados, deste modo deve-se atentar para uma cuidadosa etapa de aquisição dos dados e buscar manter a consistência dos dados durante todo o decorrer do trabalho.

Os arquivos do repositório estão organizados nas categorias eleitorado, candidatos, resultados e prestação de contas, como mostra a Figura 3.1. As subseções seguintes buscam detalhar o conteúdo de cada uma dessas categorias.

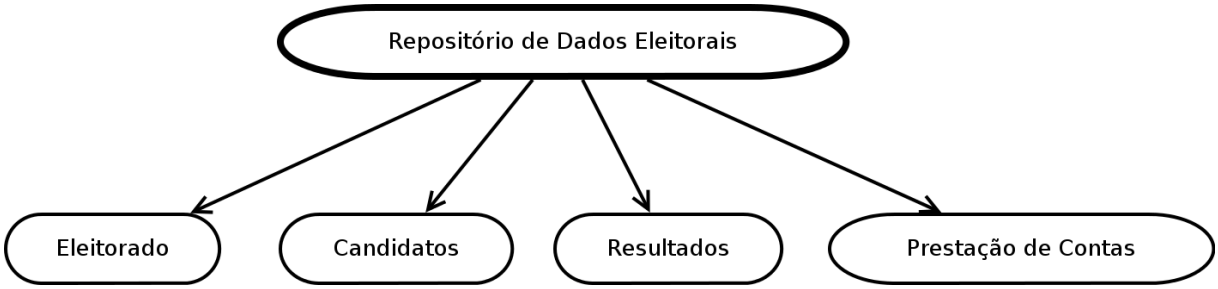

Figura 3.1: Visão geral do repositório de dados eleitorais.

3.2.1 Eleitorado

A categoria "eleitorado"contém arquivos com informações do perfil do eleitorado de cada eleição até o grão de zona eleitoral. Um município pode conter várias zonas eleitorais e o que este arquivo descreve é a quantidade de eleitores de um determinado perfil que pertencem à estas zonas.

Um perfil de eleitorado é definido pelos atributos sexo, faixa etária e grau de escolaridade. Deve-se ter em mente que o atributo grau escolaridade é declarado pelo eleitor, sendo que em muitos casos o eleitor se cadastra com um determinado grau de escolaridade, por exemplo: com ensino médio aos 16 anos de idade, e não volta ao cartório eleitoral para atualizar esta informação.

3.2.2 Candidatos

Os dados pertencentes à categoria "candidatos"são subdivididos em informações de candidatura, bens dos candidatos, legendas políticas e vagas disponíveis, como mostra a Figura 3.2.

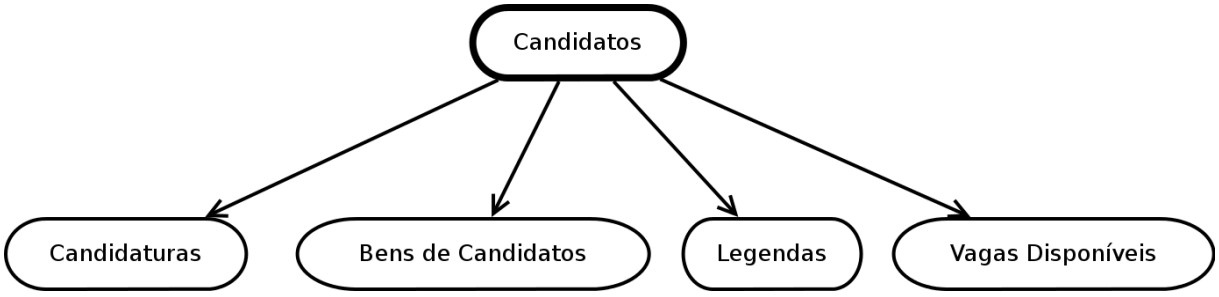

Figura 3.2: Organização dos arquivos de candidatos.

Os arquivos correspondentes às informações de candidatura são os mais extensos e também os mais importantes para as análises propostas por este trabalho. Eles contêm informações importantes sobre os candidatos, como por exemplo ocupação, sexo, idade, grau de escolaridade, estado civil e um dado crucial: se o candidato foi eleito ou não ao término da eleição. Além disso, este arquivo contém informações sobre o cargo que o candidato está concorrendo e a qual partido político e coligação ele pertence.

No momento da candidatura os candidatos são obrigados a preencher um formulário de declaração de bens e valores em cumprimento ao Artigo 13 da Lei n^o 8.429 [4]. O repositório de dados do TSE disponibiliza os conteúdos desses formulários, para cada item declarado por um candidato há uma descrição, o valor e o tipo do item.

De acordo com o arquivo de ajuda do repositório de dados do TSE, uma legenda política pode ser constituída por um ou mais partidos políticos. Além disso, diferentes legendas podem ser formadas em diferentes unidades eleitorais do país e também para diferentes cargos.

Os arquivos de vagas disponíveis são os mais simples da categoria "candidatos"e descrevem para cada unidade eleitoral a quantidade de vagas disponíveis para cada cargo.

3.2.3 Resultados

O repositório de dados disponibiliza os resultados das eleições por meio de duas categorias: votos totalizados e boletins de urnas. A organização completa desses arquivos pode ser observada

na Figura 3.3.

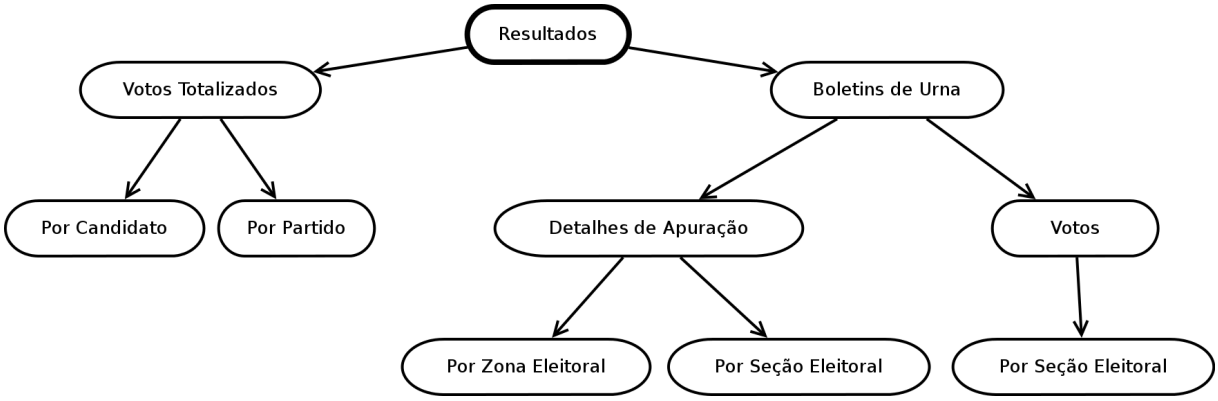

Figura 3.3: Organização dos arquivos de resultados de eleições.

Os votos totalizados são o resultado do trabalho realizado pelo sistema de apuração, ou sistema de totalização. Para cada candidato é informado o número de votos obtidos em cada zona eleitoral de cada município. A mesma informação é disponibilizada para cada partido político.

Os boletins de urna são os dados coletados pelas seções eleitorais. Pode-se ainda dividir esses boletins em duas categorias: detalhes de apuração e votos. Os detalhes de apuração descrevem para zonas eleitorais e para seções eleitorais uma série de informações em relação a uma determinada eleição. Esses detalhes informam o comparecimento e as abstenções dos eleitores e o número de votos válidos, nulos, em branco e em legenda. Por meio dos boletins de votos realiza-se a apuração de uma eleição. Este tipo de boletim revela quantos votos um candidato de um determinado partido obteve em cada seção eleitoral.

3.2.4 Prestação de Contas

Os arquivos de prestação de contas descrevem as receitas e despesas de campanha dos candidatos, partidos e comitês. As prestações de contas estão disponíveis em meio digital somente a partir de 2002.

O TSE não faz menção aos arquivos de prestação de contas em seu arquivo de ajuda. Existe, na verdade, para cada eleição um arquivo específico que descreve como a informação de prestação contas se encontra para aquela eleição particularmente. A falta de uma especificação geral para esses dados resulta em uma maior complexidade no tratamento desses dados, pois a ordem e, em alguns casos, até mesmo a existência de determinados campos varia de eleição para eleição.

De um modo geral, o repositório de dados disponibiliza os valores e pessoas (ou empresas) envolvidos em cada receita e despesa dos candidatos, partidos e comitês. Para doadores e fornecedores têm-se por exemplo os seus nomes e os números do CPF, ou CNPJ tratando-se de uma empresa.

3.3 Seleção dos Dados

O levantamento dos dados permite definir quais dados são relevantes para as análises propostas. A tarefa de seleção dos dados também é necessária, pois é impossível trabalhar com todos os dados do repositório de dados dentro do cronograma estipulado. Assim, com o intuito de manter o trabalho dentro do escopo definido, selecionou-se apenas dados das categorias "candidatos"e "prestação de contas".

Dentro da categoria "candidatos"escolheu-se utilizar as informações de vagas disponíveis, candidaturas e bens de candidatos, pois tratam-se dos dados mais diretamente relacionados às análises propostas. Para a categoria "prestação de contas"apenas os gastos e receitas de candidatos foram considerados.

Verificou-se que para eleições anteriores à 2006 algumas informações essenciais, como número do título de eleitor do candidato, não constavam para todas entradas. Então, decidiu-se por trabalhar somente com os dados de 2006 até 2012.

3.4 Criação do Banco de Dados

A fase de criação do banco de dados consiste em interpretar os dados selecionados de modo que seja possível realizar consultas para analisar relações entre os dados, como por exemplo: informações de candidaturas de um determinado candidato ao longo de diferentes eleições.

O banco de dados se comporta como um centralizador de toda a informação adquirida esparsamente no repositório de dados. Em sua modelagem e implementação deve se ter em mente que a síntese de informações de diferentes arquivos requer um grande esforço, assim à esta etapa reservou-se uma grande parcela do tempo de execução deste trabalho.

Optou-se por dividir a etapa de criação do banco de dados em módulos que são detalhados nas subseções seguintes. Assim tanto a implementação quanto a descrição do trabalho realizado podem ser melhor compreendidos.

3.4.1 Tratamento de Ambiguidade entre Palavras

3.4.2 Dados de Localidade

Uma informação que é recorrente em praticamente em todos os arquivos do repositório de dados trata-se da localização de determinado acontecimento, por exemplo o local onde um candidato concorre a um determinado cargo, ou o seu local de nascimento. Essa informação representa um dado valioso para as análises propostas, pois sua disponibilidade viabiliza análises sobre tendências geográficas.

No entanto existe um problema no tratamento dessa informação, o repositório do TSE utiliza um código gerado internamente para identificar os municípios. Esse código não apresenta relação com outros sistemas do governo, assim no caso de análises geográficas não é possível, por exemplo, utilizar dados do IBGE para determinar a localização de um município.

A solução para contornar essa questão foi encontrada em um recurso criado pelo Departamento de Informática do SUS, o DATASUS [5]. Este recurso é um trabalho que tem como objetivo o aprimoramento da gestão do sistema de saúde nacional. Apesar de se tratar de um recurso voltado para a saúde, esse sistema apresenta uma rica base de dados com arquivos de diversas categorias, incluindo uma categoria de unidades territoriais.

As tabelas do DATASUS relacionadas à unidades territoriais apresentam uma hierarquia completa de níveis de detalhamento geográfico que possibilitam trabalhar desde regiões nacionais até distritos de um município, como pode ser observado na Figura 3.4.

Para cada município tem-se sua localização geográfica com base em sua latitude e longitude. Além disso, o sistema mantém códigos em suas tabelas que possibilitam a comunicação com outros sistemas do governo, inclusive com o IBGE.

A proposta deste modulo é processar os dados do SUS e permitir que em processamentos futuros seja possível estabelecer uma relação entre os dados de localidade encontrados no repositório

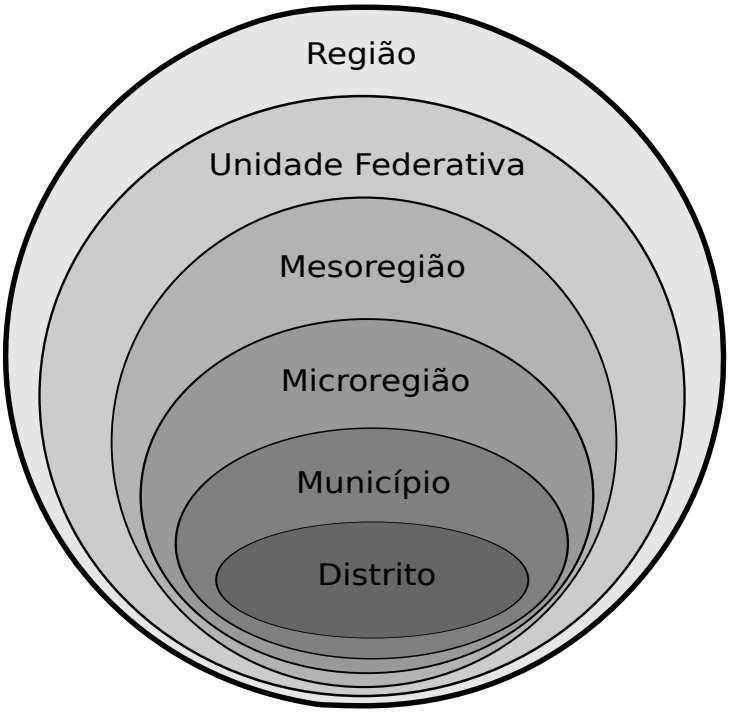

Figura 3.4: Hierarquia dos níveis de detalhamento geográfico dos dados de unidades territoriais do DATASUS.

do TSE e os dados do SUS. Para estabelecer essa relação bastaria utilizar o nome da cidade e seu estado, já que o nome da cidade é único para um estado.

Os dados do DATASUS estão no formato dBase e podem ser baixados livremente e importados por diversos softwares que suportam este formato. Os arquivos se encontram na codificação DOS/850 e esta deve ser levada em conta ao se realizar a importação dos arquivos. Este trabalho utilizou a ferramenta livre LibreOffice [6] para converter os arquivos dBase para o formato CSV e assim posteriormente processá-los juntamente com os dados do TSE.

Desenvolveu-se um processador responsável por carregar os arquivos sobre unidades territoriais do DATASUS no banco de dados. Esta tarefa resume-se a interpretar cada campo do arquivo CSV em um atributo de uma nova tabela no banco de dados. O resultado deste processamento é apresentado na Figura 3.5.

A tabela de regiões contém 6 entradas (*rows*), 1 que designa região ignorada ou no exterior e 5 que representam as macrorregiões nacionais. Uma região contém um ou mais estados.

Ao todo a tabela de estados contém 28 entradas, sendo elas:

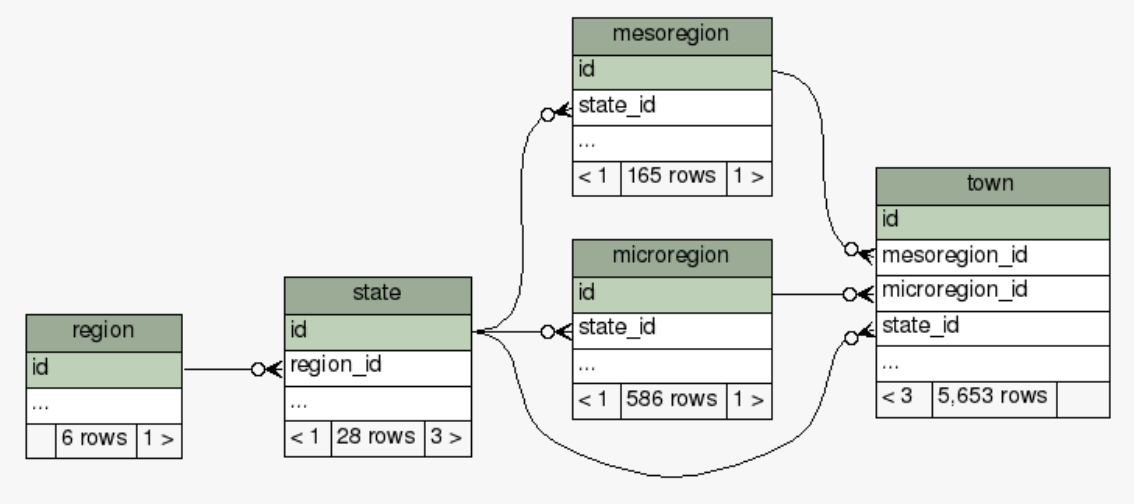

Figura 3.5: Resultado do processamento dos dados de unidades territoriais. Diagrama construído pela ferramenta *SchemaSpy* [7]

- 26 Unidades Federativas;
- Distrito Federal; e
- UF ignorada ou no exterior.

As tabelas de mesorregiões e microrregiões com 165 e 586 entradas, respectivamente, representam subdivisões dos estados que congregam diversos municípios de uma área geográfica com similaridades econômicas e sociais. Ambas foram criadas pelo IBGE para fins estatísticos e não constituem entidades políticas ou administrativas.

A tabela de cidades contém informações referentes a:

- 5.565 Municípios em vigor, sendo 5.563 Municípios propriamente ditos, 1 Distrito Federal (Brasília) e 1 Distrito Estadual (Fernando de Noronha);
- 60 Municípios transferidos de Goiás para Tocantins, quando da criação deste; estes códigos são válidos até 31/12/1988;
- 1 Município extinto (Pinto Bandeira, no Rio Grande do Sul, reincorporado a Bento Gonçalves por decisão judicial); e
- 27 Municípios ignorados, um para cada Unidade da Federação (exceto do Distrito Federal) e um correspondendo a UF ignorada ou no exterior.

Os demais níveis de detalhamento, como distritos de municípios, não foram processados pois estão fora do interesse deste trabalho.

O processamento descrito nesta subseção estabelece uma base confiável de informações sobre as unidades territoriais do Brasil e servirá como referência no tratamento de todos atributos de localidade nos processamentos descritos nas subseções seguintes.

3.4.3 Dados de Vagas Disponíveis

Os arquivos de vagas disponíveis contém informações sobre todas as eleições realizadas em todas unidades eleitorais. Assim, este arquivo possibilita não somente registrar o número de vagas para cada cargo como também criar a tabela eleição como mostra a Tabela 3.1.

Atributos da Tabela Eleição
Ano
Descrição
Estado
Cidade
Código do Cargo
Descrição do Cargo
Número de Vagas

Tabela 3.1: Atributos da tabela de eleições, resultante do processamento dos arquivos de vagas disponíveis.

Durante o processamento dos arquivos de vagas detectou-se situações em que não havia uma correspondência entre os nomes de alguns municípios apresentados nos arquivos do TSE e os dados obtidos do SUS. Após de uma análise minuciosa das ocorrências deste problema, encontrou-se os seguintes motivos e soluções para essas situações:

- Em 24 dos casos o problema é causado por divergências na grafia dos nomes como em: “Mogi Mirim” e “Moji Mirim”. Situações semelhantes ocorrem pelo uso do apostrofo em nomes como “Santa Luzia D’Oeste”. O problema foi tratado pela abordagem descrita na Subseção 3.4.1 e determinou-se a correspondência entre os nomes para todos os 21 casos.

- 8 dos casos justificam-se por alterações nos nomes dos municípios, como por exemplo o município de “Januário Cicco” no Rio Grande do Norte que passou a se chamar “Boa

Saúde". A solução adotada foi buscar para cada caso o nome antigo do município nos registros do IBGE [8], conseguindo assim estabelecer a correspondência entre os nomes para esses casos.

- Para 5 casos verificou-se que os municípios ainda não haviam sido instalados quando os dados do SUS/IBGE foram coletados. Por isso não constam na base de dados, tratam-se de casos como o município "Paraíso das Águas" no Mato Grosso do Sul. A única solução para esses casos foi armazená-los na base de dados mesmo sem uma referência com os dados do SUS/IBGE. Esses municípios não poderão ser utilizados em todas as análises propostas.

O processamento descrito nessa seção foi responsável por armazenar 28.082 eleições no banco de dados.

O processamento dos dados utilizando o tratamento do nome das palavras resultou em um total de 28.082 eleições em 5.568 cidades. Processou-se os arquivos pertencentes ao período de 2006 à 2012, resultando em um total de 28.082 eleições em 5.568 cidades. Para 5 cidades não foi possível realizar a correspondência com os dados do SUS mesmo após à etapa de tratamento de ambiguidade mencionado na Subseção 3.4.1. Mesmo essas cidades foram armazenadas no banco de dados, no entanto não poderão ser utilizadas em todas as análises propostas.

3.4.4 Dados de Candidatura

3.4.5 Dados de Bens de Candidatos

3.4.6 Dados de Prestação de Contas

Capítulo 4

Resultados

Capítulo 5

Conclusão

Bibliografia

[1] TSE. *Repositório de Dados Eleitorais*. Ago. de 2012. URL: http://www.tse.jus.br/eleicoes/repositorio-de-dados-eleitorais.

[2] TSE. *Portal Eletrônico do Tribunal Superior Eleitoral*. Ago. de 2012. URL: http://www.tse.jus.br/.

[3] Wikipedia. *Comma-separated values*. Ago. de 2012. URL: http://en.wikipedia.org/wiki/Comma-separated_values.

[4] Brasil. *Lei número 8.429*. http://www.planalto.gov.br/ccivil_03/leis/l8429.htm.

[5] SUS. *DATASUS*. Ago. de 2012. URL: http://www2.datasus.gov.br/DATASUS/index.php?area=01.

[6] L. F. Software. *LibreOffice*. Ago. de 2012. URL: http://www.libreoffice.org/.

[7] J. Currier. *SchemaSpy*. Ago. de 2012. URL: http://schemaspy.sourceforge.net/.

[8] IBGE. *Arquivos do IBGE*. Ago. de 2012. URL: ftp://geoftp.ibge.gov.br/.