Uma abordagem NoSQL no auxílio a unificação e estudo dos registros de pessoas desaparecidas no Brasil

Raul Sena Ferreira, Alexandre de Assis Bento Lima

[1]PESC/COPPE – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – P.O. Box: 68511 – Brazil

{raulsf, assis}@cos.ufrj.br

Abstract. *According to estimates from the Federal Government, 200,000 people disappear every year in Brazil, that is, every year an equivalent population the city of Sobral, fifth most populated city in the state of Ceará, disappear from their homes without a trace. In 2002, the National Network for the Identification and Location of Missing Children and Adolescents (ReDESAP) established some goals, such as: Create a unified registration data and reported cases; Promote information sharing on the phenomenon among the various actors of the network and coordinate a collective effort nationwide. This work aims to contribute in these three guidelines, providing a model of data unification of missing persons, promoting access of innovative way for visualizing data and potencializing the systems efforts of missing persons nationwide.*

Resumo. *Segundo estimativa do Governo Federal, 200 mil pessoas desaparecem todo ano no Brasil, ou seja, a cada ano uma população equivalente a cidade de Sobral, quinto município mais povoado do estado do Ceará, some de seus lares sem deixar vestígios. Em 2002, a Rede Nacional de Identificação e Localização de Crianças e Adolescentes Desaparecidos (ReDESAP) estabeleceu alguns objetivos, dentre eles: Criar um cadastro unificado de dados e casos notificados; Promover o compartilhamento de informações sobre o fenômeno entre os diversos atores da Rede e coordenar um esforço coletivo em âmbito nacional. Este trabalho visa contribuir nessas três diretrizes, provendo um modelo de unificação nacional dos dados de pessoas desaparecidas, promovendo o acesso a informação de forma inovadora e potencializando os esforços de sistemas de cadastro de pessoas desaparecidas em âmbito nacional.*

1. Introdução

Segundo estimativa do Governo Federal, 200 mil pessoas desaparecem todo ano no Brasil, onde 40 mil são crianças. Acredita-se que o número seja ainda maior, pois segundo um estudo feito por [de Oliveira 2007], inúmeros casos não são registrados e nem investigados por vários motivos adversos e além disso, ainda segundo o estudo, o conceito de desaparecimento é juridicamente ligado ao conceito de falecimento, o que pode tornar a busca menos interessante para os órgãos responsáveis.

Segundo o SEDH(Secretaria Especial dos Direitos Humanos), cerca de 15% dos desaparecimentos não são resolvidos por um longo período de tempo ou jamais são solucionados. Pensando nisso, em 2008 aconteceu uma cuidadosa atualização e revisão do Programa Nacional de Direitos Humanos I e II, tendo como instrumento fundamental a

realização da 11ª Conferência Nacional dos Direitos Humanos – 11ª CNDH [1] culminando na PNDH III.

A Secretaria de Direitos Humanos da Presidência da República – SDH/PR, por meio da Secretaria Nacional de Promoção dos Direitos da Criança e do Adolescente - SNPDCA formalizou em 2002 a Rede Nacional de Identificação e Localização de Crianças e Adolescentes Desaparecidos – ReDESAP. A ReDESAP é composta pelo Conselho Nacional dos Direitos da Criança e do Adolescente - CONANDA, Fórum Colegiado Nacional dos Conselheiros Tutelares, Representantes de entidades não governamentais de apoio e atendimento as famílias de crianças e adolescentes desaparecidos, representantes das Secretarias de Segurança Pública dos estados, órgãos e entidades públicas e privadas, agencias e organismos internacionais, universidades. [2]

Objetivos da Rede Nacional de Identificação e Localização de Crianças e Adolescentes Desaparecidos (ReDESAP):

- Criar um cadastro unificado de dados e casos notificados
- Articular serviços especializados de atendimento público
- Promover o compartilhamento de informações sobre o fenômeno entre os diversos atores da Rede
- Coordenar um esforço coletivo em âmbito nacional

Porém apesar de todo o esforço empregado existem alguns problemas relevantes que precisam ser dirimidos como:

- Muitos sites com dados de adultos e crianças desaparecidas, buscar em qual?
- Muita informação repetida, desatualizada e incompleta
- Cadastro Nacional de Crianças e Adolescentes Desaparecidos: 370 casos registrados, provavelmente está muito abaixo da realidade
- Estatísticas oficiais desatualizadas e não condizentes com a realidade

Além disso, os dados que existem hoje não estão disponibilizados de forma organizada, já que não existe uma interface de consulta que permita extrair maior informação sobre os dados, a componente geográfica dos dados não é explorada, além disso, não existe uma interface que permita que as pessoas possam acessar e analisar, de forma simples, os dados.

Por isso, este trabalho apresentará um modelo de solução paliativa para os problemas mencionados acima, dividindo-se assim em três partes, a saber:

1. Modelo de unificação das diferentes bases de dados existentes
2. Modelo de interface de consulta e análise dos dados já coletados
3. API disponibilizando publicamente os dados já coletados

O trabalho é organizado da seguinte forma: A Seção 2 mostra alguns trabalhos relacionados na área, a Seção 3 descreve o problema e a proposta deste trabalho. Já a Seção 4 mostra a metodologia abordada, os experimentos realizados e seus respectivos resultados. E na última seção apresentamos as conclusões extraídas de todo o trabalho realizado bem como os próximos passos a serem tomados.

[1] Disponivel em: http://www.ipea.gov.br/participacao/images/pdfs/conferencias/Direitos_humanos_XI/relatorio_regulamento_

[2] Mais detalhes: http://www.desaparecidos.gov.br/index.php/redesap

2. Trabalhos Relacionados

Alguns sistemas foram criados na esperança de tentar minimizar o problema do déficit no registro de pessoas desaparecidas em relação a real quantidade de casos de desaparecimento, um deles é o Cadastro Nacional de Pessoas Desaparecidas CNPD que é mantido pela Secretaria de Direitos Humanos da Presidência da República (SDH/PR) com o apoio do Ministério da Justiça (MJ).

Este sistema por sua vez, tem diversos desafios a enfrentar e um dos principais é a composição e manutenção de estatísticas nacionais a respeito do tema. Para consolidar uma matriz nacional de informações a esse respeito, o sistema tenta mapear iniciativas estaduais de registro e divulgação de casos de pessoas desaparecidos e com o apoio das redes de segurança pública e de direitos da criança e do adolescente, registrá-los na base nacional.

Porém, devido a dificuldade do sistema se manter atualizado outros sistemas foram nascendo de forma independente e criando suas próprias bases de dados, permitindo que as pessoas possam cadastrar seus entes ou amigos desaparecidos em vários sites. Alguns desses sistemas conseguem inclusive, compartilhar informações do desaparecido através das mídias sociais, como é o caso do aplicativo BiaMap.

Existem também diversos outros sites mantidos pelas delegacias civis de vários estados do Brasil, cada um com sua própria base de dados, enquanto vários outros estados não possuem um sistema disponível na internet com o mesmo fim.

Outras iniciativas também foram tomadas, como o projeto ”Caminho de Volta” [Gattás et al. 2007] onde um dos pontos de destaque do programa se dá através da implantação de um banco de DNA, onde as famílias cedem uma gota de sangue para a análise do perfil do DNA e o material fica arquivado com segurança, para fins de posterior identificação. Já [Figueredo and Rodrigues de Souza 2013] tenta ajudar a descobrir pessoas desaparecidas a partir do reconhecimento facial automatizado, aproveitando-se da crescente quantidade de câmeras que são utilizados em espaços públicos.

O site da patrulha do Missouri, um estado norte americano, desenvolveu um trabalho com um viés mais geográfico ao mostrar em um mapa do território do estado, as quantidade de pessoas desaparecidas separadas por cidade e suas respectivas informações. Enquanto no Brasil, foi feito um mapeamento dos desaparecimentos também de forma geográfica, cruzando os dados de desaparecimentos, autos de resistência e homicídios, fornecidos pelas delegacias de polícia do estado do Rio de Janeiro, onde apesar de ter apenas os pontos em nível estadual, já foi possível descobrir dados interessantes, dentre elas que a zona sul, a região mais rica da cidade, tem a menor taxa de pessoas não localizadas, enquanto zona oeste lidera o ranking das pessoas ainda não encontradas.

3. Motivação e Objetivo

A informação geográfica tem grande importância em diversas áreas como, marketing, agricultura, meio ambiente, saúde, planejamento urbano entre outros, ajudando na tomada de decisões e estratégias, agregando a análise um meio de representação visual mais expressiva do que uma representação discreta. Várias ferramentas podem ser criadas visando extrair o máximo de informações agregadas a distribuição espacial, informações essas que não poderiam ser extraídas através do modo convencional de análise de dados

não espacial.

A informação geográfica, já existe há centenas de anos, (e.g., mapas) e como em vários aspectos do nosso cotidiano esta também foi e vem sendo alterada pela modernidade tecnológica. Para lidar com esse tipo de informação, surgiram então os sistemas de informação geográfica (SIG). Os SIG são sistemas utilizados para armazenar, analisar, manter e manipular dados geográficos de maneira automatizada.[Bolstad 2005]

Os dados geográficos utilizados pelos SIG podem ser imagens digitalizadas (e.g., fotos de satélite) ou objetos que representam uma geometria no espaço, chamados objetos espaciais. Esses dados são armazenados e gerenciados por bancos de dados espaciais (objetos geométricos espaciais).[Güting 1994]

Este trabalho tenta utilizar as vantagens do modelo de visualização geográfica em cima de dados de pessoas desaparecidas e assim propor um modelo ainda não explorado ou pouco explorado no Brasil, além de tentar trazer essas informações de forma agregada e de fácil visualização.

3.1. Definição do problema

O problema do desaparecimento de pessoas no Brasil é alarmante e ainda falta muito a fazer no quesito tecnologia, pois como foi mostrado nos capítulos anteriores, ainda não possuímos um sistema de banco de dados nacional e integrado. Junto a isso, soma-se o fato de não existir em todos os estados, uma delegacia especializada em desaparecimento de pessoas adultas, o que faz com que as investigações deste problema fiquem divididas com as investigações de outros problemas como casos de homicídios, roubos, sequestros, entre outros crimes.

Por outro lado, apesar da existência de delegacias especializadas em desaparecimento de crianças e adolescentes, este ainda dispõe de um sistema que sofre com a defasagem de atualização, integração com outros bases e não disponibiliza de forma fácil os dados, para que a sociedade como um todo possa estudar e observar esses dados, bem como órgãos governamentais e não governamentais possam cruzar esses dados com outros dados provindos de outras fontes como, de indigentes, de homicídios, e etc.

Várias iniciativas foram tomadas mas estes sistemas acabam se tornando ilhas de informação, onde se obtém a vantagem de se poder cadastrar informação de forma mais veloz do que os órgãos públicos mas que acabam por vezes obtendo os registros de uma parcela das pessoas desaparecidas, pois o registro nestes sites fica condicionado a quantidade de pessoas que conhecem o site. Desta forma temos vários sites com registros mas nenhum sistema que tenha todos esses dados em um lugar só.

Sendo assim, não existe atualmente no Brasil, uma base de dados unificada, além disso, as informações de localização ou desaparecimento não são exploradas, ou seja, os dados além de espalhados por diversas fontes, muitas vezes repetidos, também estão incompletos, dependendo da fonte, outro fator importante, é que os dados não são georreferenciados, ou seja, estamos perdendo a chance de estudar e visualizar os dados de forma mais rica, aproveitando todas as vantagens que uma representação geográfica pode fornecer.

3.2. Proposta

A proposta concentra-se em desenvolver um sistema que se divide em três frentes:

1) Desenvolver um modelo que unifique os dados existentes nas grandes bases de dados de pessoas desaparecidas do país, ou seja, os sites de pessoas desaparecidas construídos por reconhecidas ONGs, empresas que já trabalham em sistemas de cadastro de pessoas desaparecidas e dos sites de órgãos públicos como os da policia civil ou da secretaria de direitos humanos da presidência da república. Desta forma espera-se que o sistema consiga minimizar o problema da descentralização dos registros e ainda consiga potencializar o que os demais sites fazem de melhor: registrar e divulgar de forma rápida os dados de um desaparecido (com seu B.O. devidamente preenchido).

2) Desenvolver um modelo de interface de consulta que permita ao usuário do sistema visualizar e extrair informação de valor dos dados através de um mapa, dando ao usuário uma percepção do ponto de vista geográfico, que até então não existe. Outra funcionalidade desejada é que o sistema permita que gráficos sejam gerados a partir da consulta, visando permitir que o usuário do sistema entenda melhor as diversas informações que o dado traz consigo, como, estudo de questões étnicas, faixa etária, entre outras condições que possam permitir uma percepção mais profunda do problema. Além disso, o sistema deve mostrar os dados de forma agregada, mantendo sempre as fontes de onde os dados foram retirados, enriquecendo assim cada registro recuperado, tornando mais valiosa cada informação inserida por um tipo de fonte.

3) Desenvolver uma API, de acesso público, que disponibilize os dados coletados para quem quiser obtê-los, para que outras pessoas, pesquisadores ou desenvolvedores independentes, possam estudar os dados e construir novas aplicações que possam ajudar a entender melhor o problema e/ou propor novas soluções. Tornar os dados públicos e de fácil acesso pode ajudar a própria sociedade a se envolver mais com o tema e assim, tornar a causa mais conhecida e promover a novas soluções em cima desses dados.

A figura 1 ilustra o modelo usado neste trabalho, onde basicamente o sistema captura informações de diferentes sites de pessoas desaparecidas, e armazena esses dados em um banco de dados orientado a documentos[Han et al. 2011], no intuito de facilitar a gravação dos dados sem alterar suas estruturas originais, que por sua vez, são heterogêneas. Em seguida Um processo de ETL (*Extract, Transform and Load*) lê os dados do banco, extrai todas as informações possíveis e carrega de forma estruturada para um banco relacional, para que as consultas convencionais e espaciais possam ser processadas e enviadas para a página da consulta.

Depois que a consulta é realizada e antes de ser enviada para o navegador, o sistema calcula as funções de probabilidade de densidade, através do método de estimador de kernel (KDE), que será melhor explicado no próximo capítulo, e aí então, o resultado da consulta e do processamento estatístico é enviado ao navegador em formato de grupos de pontos e gráficos. No meio do caminho, existe um banco de dados chave-valor que trabalha como um cache da consulta, diminuindo a latência de consultas que são realizadas com frequência, fazendo com que o cálculo do KDE juntamente com a consulta dos dados só seja feita uma vez a cada nova configuração de consulta, aumentando a velocidade na resposta da aplicação.

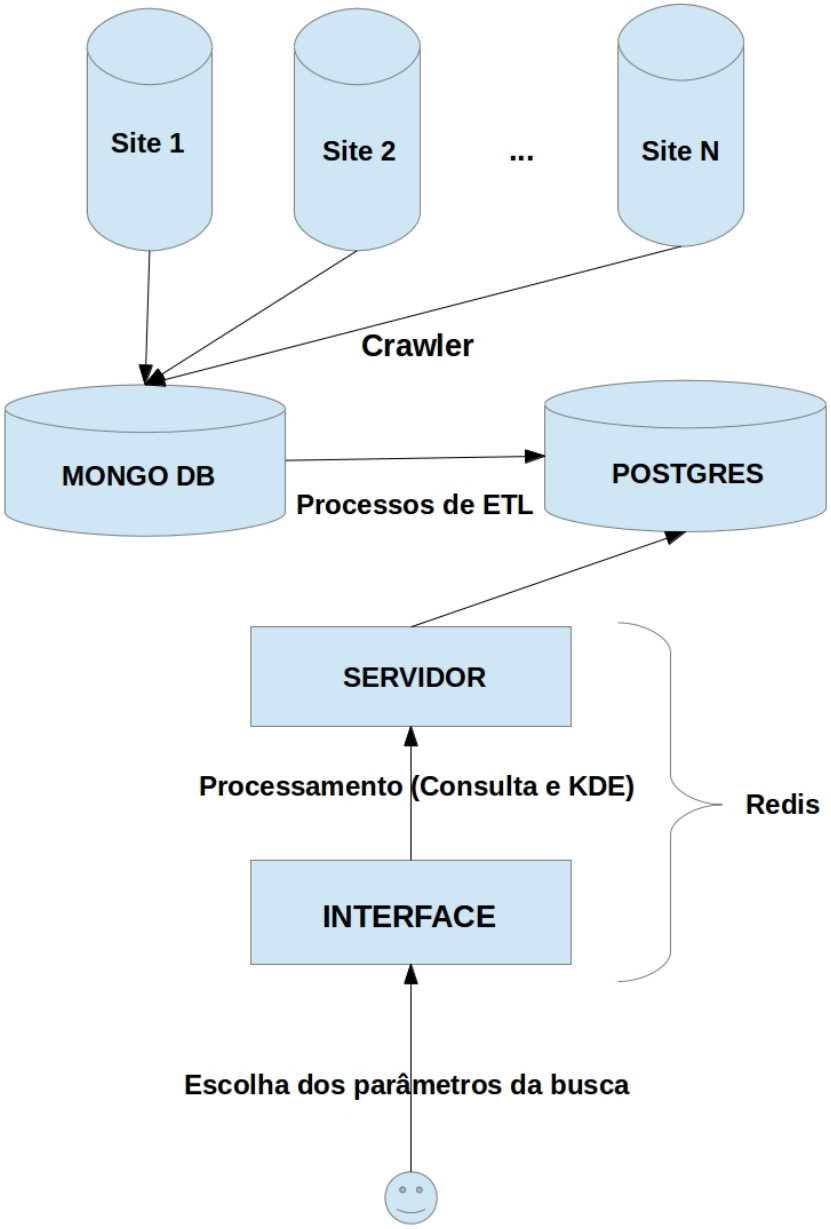

Figure 1. Arquitetura da proposta

4. Experimentos

Abaixo uma breve explicação sobre os elementos utilizados no experimento bem como os resultados obtidos nos primeiros testes do sistema. Os experimentos foram realizados com o sistema hospedado em um servidor na nuvem, onde o mesmo pode ser acessado através do link: http://www.projeto-myosotis.com.br

4.1. Estimando Densidades

Este sistema fará uso de um método estimador de densidade, para tentar estimar a probabilidade de uma nova ocorrência baseada na localização das ocorrências e na quantidade

de vezes que elas aparecem em um estado. Embora a precisão deste método fique condicionada ao fato de obter-se todas as ocorrências de todos os estados, ainda assim, pode ser interessante saber e visualizar em um mapa, o quão grande é a quantidade de desaparecimentos em determinados estados e como essa informação se relaciona com as demais informações textuais contidos nos dados.

Para este trabalho, foi escolhido o mais conhecido estimador de densidades, no caso, o estimador de densidade de kernel (ou Kernel Density Estimation), também conhecido como Janela de Parzen [Duda et al. 2012]. Neste método são utilizadas funções não-lineares como Gaussianas e Sigmóides para se computar a densidade local de cada instância. A densidade de uma população pode ser estimada com várias técnicas estatísticas, porém estatisticamente, alguns dados ou populações não possuem estruturas ou parâmetros característicos, no caso, estes dados são conhecidos como não paramétricos.

A função de probabilidade é um conceito fundamental em estatística e existem diversas técnicas que podem ser empregadas para estimar dados não paramétricos. Os dados usados no sistema podem ser considerados dados não paramétricos, pois não dependem de dados pertencentes a nenhuma distribuição particular. Tipicamente, o modelo não-paramétrico cresce no sentido de acomodar a complexidade dos dados. Como métodos não paramétricos fazem menos suposições, a aplicabilidade deles é mais larga que os correspondentes métodos paramétricos. Em particular, eles podem ser aplicados em situações em que menos se sabe sobre o problema em questão.

Além disso, devido a menor dependência de hipóteses, métodos não paramétricos são mais robustos. Um exemplo de dado não-paramétrico: distribuição tem a forma normal, tanto a média quanto a variância não foram especificadas.

4.2. Base de dados e ETL

A base de dados foi obtida através da extração automática dos registros de quatro fontes diferentes: O cadastro nacional de pessoas desaparecidas [3]; Portal de desaparecidos mantido pelo Ministério da Justiça [4]; Cadastro de pessoas desaparecidas da polícia civil do Rio Grande do Sul [5] e do sistema BiaMap [6]

Primeiramente, os dados eram armazenados em coleções distintas no banco de dados orientado a documento, onde cada site era tratado como uma coleção e cada registro desse site, um documento. Em seguida, de posse das quatro coleções de documentos, foi desenvolvido um processo de ETL para cada coleção de dados, limpando e extraindo informações para serem colocadas posteriormente em formato de atributos, de forma agregada, em um banco de dados relacional, resultando assim em uma base de dados unificada.

A base de dados unificada, inicialmente contém 26 atributos: id, nome, foto, sexo, cor dos olhos, cor da pele, cabelo, peso aproximado, altura aproximada, tipo físico, indicativo de transtorno mental, idade, data de nascimento, quantidade de dias desaparecido, data do desaparecimento, bairro, cidade e estado onde ocorreu o desaparecimento, marca de nascença, status (desaparecido ou encontrado), informações sobre o caso, indicativo

[3]http://www.desaparecidos.gov.br
[4]http://portal.mj.gov.br
[5]http://www.desaparecidos.rs.gov.br
[6]http://www.biamap.com.br

de registro de boletim de ocorrência, fonte da informação, latitude, longitude e um campo de geometria.

4.3. Resultados

A informação de vários registros foram beneficiados com a agregação da informação, já que vários registros em algumas fontes não tinham alguns atributos enquanto possuía em outros, além do mais, as informações de geolocalização, que não existia em nenhuma fonte, foram acrescentadas ao registro final. A figura 2 mostra a tela inicial da consulta enquanto as figuras 3 e 4 mostram, respectivamente, as telas do mapa e dos gráficos resultantes da consulta.

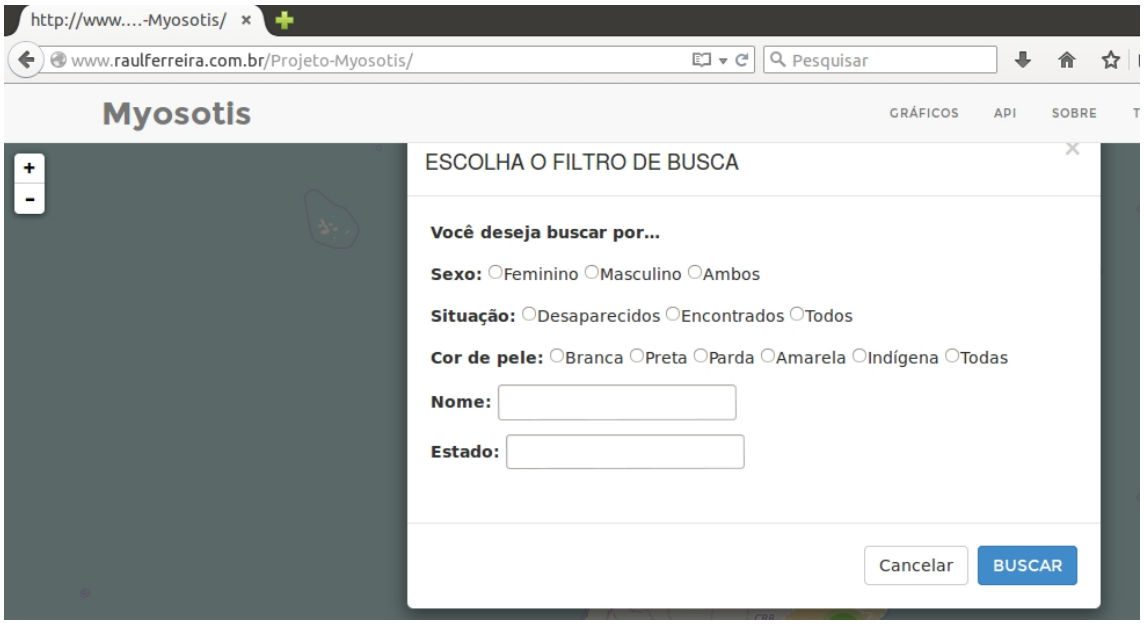

Figure 2. Interface de consulta

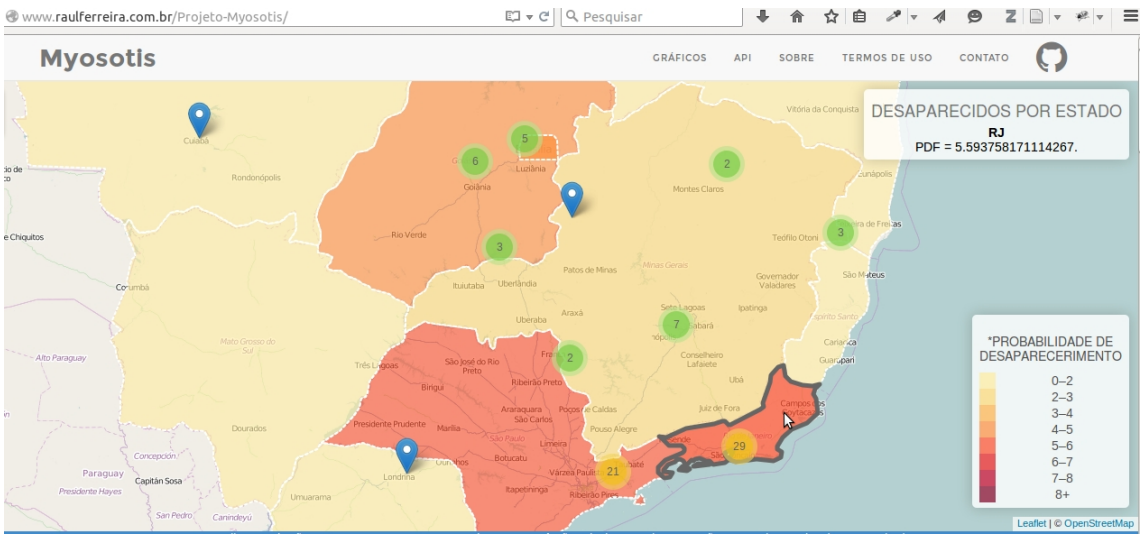

Figure 3. Resultado no mapa

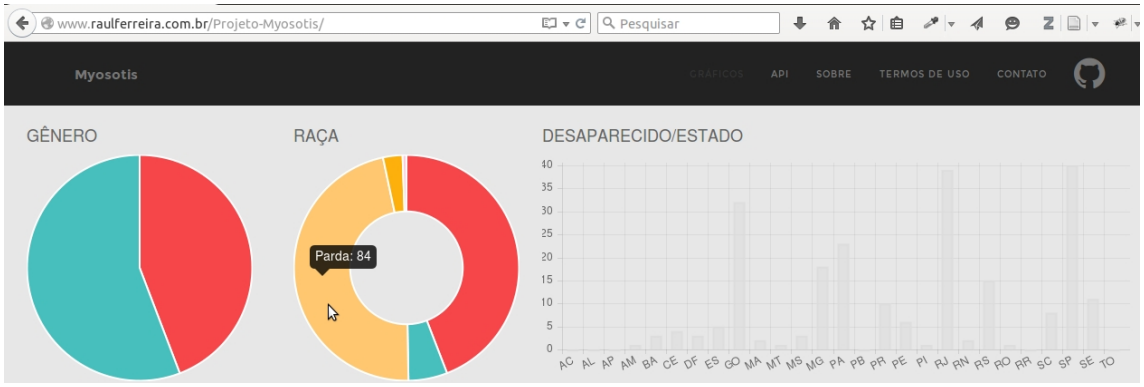

Figure 4. Gráficos resultantes

A quantidade de registros distintos no final foi maior do que nos sites buscados, como era o esperado. Em números o resultado em cima das quatro primeiras fontes foi:

- 1794 Registros distintos
- 1600 Registros com coordenadas geográficas
- 1023 Registros de pessoas ainda desaparecidas

5. Conclusão e Trabalhos Futuros

Com os experimentos realizados e uma primeira versão do sistema lançado, podemos concluir que os resultados foram satisfatórios, já que a ferramenta conseguiu fazer o que havia proposto, e de fato, os registros se tornaram mais ricos no quesito informação, tanto na parte textual quanto na parte geográfica. Além disso, o resultado da visualização no mapa de forma intuitiva criou a possibilidade de extrair informações que antes não podiam ou eram difíceis de serem percebidas analisadas apenas os dados de forma numérica ou textual.

Um ponto negativo nesta primeira versão é o fato de se ter extraído os dados de poucas fontes, mas com a adição de novos sites no sistema, este problema deverá ficar menor. Outro problema por enquanto é a fragilidade do sistema de ETL, já que este precisará sofrer alterações todas as vezes que a estrutura de algum desses sites for alterada, porém como a alteração estrutural de um sistema ou site não é algo que aconteça corriqueiramente, espera-se que este não seja um grande problema.

Ainda existe muita coisa a ser feita como trabalho futuro, porém os principais caminhos a serem tomados em um primeiro momento podem ser:

- Extrair os dados dos sites restantes, o qual ainda faltam vários
- Extrair mais informações dos dados (ex: bairros e logradouros)
- Incluir o estudo de séries históricas dos dados
- Incluir mais tipos de gráficos e cruzamento de dados de outras bases, exemplo, indigentes, homicídios e etc
- Por ser de código aberto, fomentar a contribuição à ferramenta através de comunidades e fóruns de código livre
- Tornar a visibilidade da ferramenta maior e transformá-la em um sistema reconhecidamente como um unificador de bases

References

Bolstad, P. (2005). *GIS Fundamentals: A First Text on Geographic Information Systems.* Eider Press.

de Oliveira, D. D. (2007). *Desaparecidos civis: conflitos familiares, institucionais e segurança pública*. PhD thesis, Universidade de Brasília. Brasília.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Figueredo, M. B. and Rodrigues de Souza, J. (2013). Face recognition model applied to the missing people problem. In *Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on*, pages 1–3. IEEE.

Gattás, G. J. F., Figaro-Garcia, C., Massad, E., Battistella, L. R., Fridman, C., Lopez, L. F., Wen, C. L., Neumann, M. M., Sumita, C. H., Vieira, M. R., et al. (2007). Título: Caminho de volta: Tecnologia na busca de crianças e adolescentes desaparecidos no estado de são paulo.

Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal—The International Journal on Very Large Data Bases*, 3(4):357–399.

Han, J., Haihong, E., Le, G., and Du, J. (2011). Survey on nosql database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on*, pages 363–366. IEEE.