DCluster: Um sistema para análise exploratória de grandes volumes de dados georreferenciados

Cláudio Gustavo S. Capanema[1], **Fabrício A. Silva**[1], **Thais R. M. Braga Silva**[1]

[1] Universidade Federal de Viçosa (UFV), Florestal, Brasil

{claudio.capanema, fabricio.asilva, thais.braga}@ufv.br

Resumo. *O crescimento e a diversificação do uso de dispositivos móveis, especialmente smartphones, fez surgir um grande volume de dados georreferenciados oriundos de aplicativos móveis. Como consequência, as empresas estão cada vez mais interessadas em analisar tais dados para conhecer melhor seus usuários e, assim, oferecer melhores serviços. A análise de dados georreferenciados é uma área ainda pouco explorada, e que pode trazer informações mais úteis que os dados puros. Em geral, as ferramentas atuais de análise de dados não tratam atributos georreferenciados e exigem conhecimento prévio dos usuários para a utilização de técnicas avançadas. Este trabalho propõe um sistema que visa auxiliar analistas de dados na exploração e visualização de grandes volumes de tipos variados de dados, incluindo os georreferenciados.*

1. Introdução

A utilização de dispositivos móveis está se disseminando cada vez mais rapidamente. Esse fenômeno de popularização tem trazido uma crescente geração de grandes volumes de dados oriundos da utilização de aplicativos de dispositivos como *smartphones* e *tablets*. Isso se deve a fatores como o baixo custo de aquisição e mobilidade, além do interesse dos provedores de serviços em conhecer os seus usuários. Segundo [Statista 2017], no primeiro quadrimestre de 2012, o número de usuários ativos diariamente no Facebook através de uma plataforma móvel era de aproximadamente 266 milhões. No mesmo período do ano de 2016, esse número subiu para 1,146 bilhão.

Em geral, grande parte dos dados provenientes de dispositivos móveis são georreferenciados, ou seja, incluem a localização do usuário. Um exemplo está relacionado à possibilidade de se realizar *check-ins* de localização ao interagir em redes sociais. Em outros casos, a presença do georreferenciamento é essencial para o funcionamento da ferramenta, como ocorre com os aplicativos Waze e Uber. Por fim, dados de diferentes segmentos de empresas (e.g., bancos, operadoras de telecomunicações, sistemas de comércio eletrônico, dentre outros) também possuem informações georreferenciadas. Esses dados de localização podem trazer informações relevantes para empresas e pesquisadores.

Essa crescente demanda por análise de dados fez surgir ferramentas analíticas no mercado. Entretanto, atualmente ainda existe uma deficiência no tratamento de atributos georreferenciados, que muitas vezes inexiste ou esbarra em limitantes como o preço da licença. Além disso, algumas dessas ferramentas exigem que o usuário tenha uma experiência técnica em análise de dados. Isso dificulta que pessoas com menos conhecimento técnico possam utilizar os recursos de aprendizagem de máquina, por exemplo, para trazê-los a realidade da sua área de pesquisa, o que é um aspecto que dificulta a popularização da análise de dados.

Este trabalho apresenta o DCluster, um sistema para análise exploratória de grandes volumes de dados georreferenciados. Por se tratar de uma ferramenta Web, a sua utilização não é restrita a máquinas com grande poder computacional por parte do usuário. Além disso, também é proposta uma interface simples e intuitiva, caracterizada pela ausência de informações desnecessárias, a fim de que a experiência do usuário transcorra da maneira mais amigável possível. Esses aspectos têm o objetivo de contribuir com a disseminação e a popularização da análise de dados georreferenciados.

2. Ferramentas Relacionadas

Esta seção discute as características das principais ferramentas para análise exploratória de grandes volumes de dados existentes, e as relaciona com o sistema proposto neste trabalho. Dentre os aspectos que caracterizam essas ferramentas, os mais relevantes são: suporte a dados georreferenciados, arquitetura do sistema (centralizada ou distribuída), usabilidade e licença.

Um dos sistemas mais populares é o Weka [of Waikato 2017], uma ferramenta gratuita e com interface simples. Porém, seu desenvolvimento iniciado nos anos 90, e interrompido com sua aquisição pela Pentaho em 2006, não favorece um ambiente ideal para a análise de grandes volumes de dados, uma vez que sua arquitetura é centralizada e não oferece suporte a dados georreferenciados.

O Geo-Data Visualizer [Xavier et al. 2017] é uma ferramenta gratuita desenvolvida em Javascript/Jquery que utiliza dados georreferenciados para gerar mapas e estatísticas associadas. Suas principais funcionalidades são a visualização de mapas com agrupamentos, e a filtragem de dados a partir de métricas temporais.

O BigML [BigML 2017] e o Azure Machine Learning [Microsoft 2017a] são boas alternativas para quem deseja utilizar recursos de aprendizado de máquina. Possuem interface amigável, um variado conjunto de funcionalidades, e correspondem a ferramentas Web. Ambas as ferramentas possuem planos acessíveis, porém com algumas restrições como tamanho dos dados. No entanto, ambas as ferramentas esbarram na falta de suporte a dados georreferenciados.

Os sistemas Pentaho Big Data [Hitachi 2017], Tableau [Tableau 2017], Microsoft Power BI [Microsoft 2017b], SAS Business Intelligence and Analytics [SAS 2017], Qlik [Qlik 2017] e Sisense Business Analytics Software [Sisense 2017] se enquadram no conjunto de ferramentas pagas de análise de dados mais completas da atualidade: suportam diversas fontes de dados, possuem versões online, interfaces intuitivas, e trabalham com dados georreferenciados. Essas ferramentas se enquadram na categoria de BI (*Business Intelligence*), e sua ampla lista de funcionalidades faz com que o aprendizado seja relativamente complexo. Além disso, o preço das licenças é muitas vezes inviável para pequenas e médias empresas.

3. DCluster

O DCluster consiste em um sistema Web para análise exploratória de grandes volumes de dados, com foco no georreferenciamento. Sua arquitetura cliente-servidor

Tabela 1. Comparação entre sistemas.

Parâmetros / Sistemas	Georreferenciamento	Licença	Plataforma
Weka	ausente	gratuita	desktop
Geo-Data Visualizer	presente	gratuita	web
BigML	ausente	paga*	web
Azure Machine Learning	ausente	paga	web
Pentaho Big Data	presente	paga	web e desktop
Power BI	presente	paga	web e desktop
Tableau	presente	paga	web e desktop
SAS	presente	paga	web
Qlik	presente	paga	web e desktop
Sisense	presente	paga	web e desktop

tem por finalidade facilitar a utilização da ferramenta, uma vez que não é necessária a instalação em máquina local. O DCluster também se destaca pela utilização de recursos interativos para a visualização de gráficos, e por ser desenvolvido na linguagem Python, que oferece uma lista de APIs para análise, visualização de dados e aprendizagem de máquina. O sistema apresenta um conjunto de funcionalidades para processar, visualizar e exportar dados. A Figura 1 ilustra os principais componentes do DCluster.

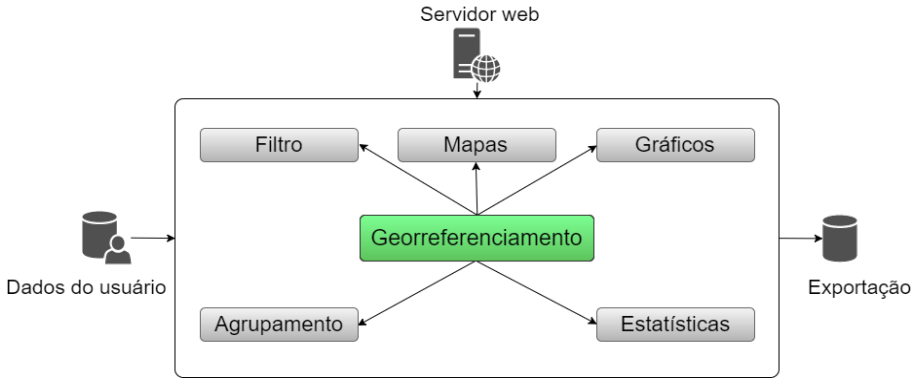
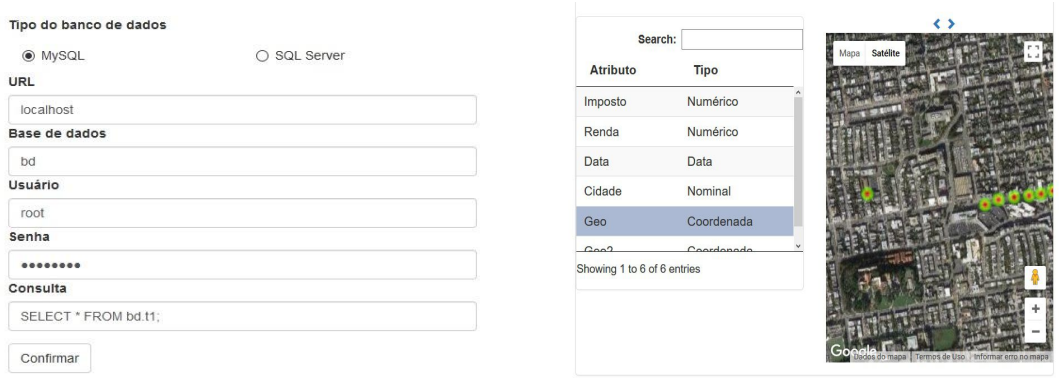


Figura 1. Diagrama de funcionalidades

3.1. Entrada de Dados

O fluxo de execução se inicia com o usuário enviando ao servidor os dados tabulares a serem processados (veja a Figura 2(a)), que podem ser provenientes de um arquivo no formato CSV (*Comma Separated Values*), ou pela conexão a um banco de dados MySQL ou SQL Server. O sistema então identifica automaticamente os tipos de cada atributo (coluna) dos dados enviados, que podem ser: numérico, nominal ou data/hora. Vale destacar que os atributos georreferenciados são difíceis de serem identificados automaticamente, por se tratarem em geral de valores numéricos reais representando a latitude e longitude. Com isso, os atributos de latitude e longitude são inicialmente assumidos como numéricos. A indicação de quais atributos representam a latitude e longitude de uma localização deve ser feita manualmente pelo usuário.



(a) Entrada de dados (b) Visualização de mapa

Figura 2. Telas do DCluster

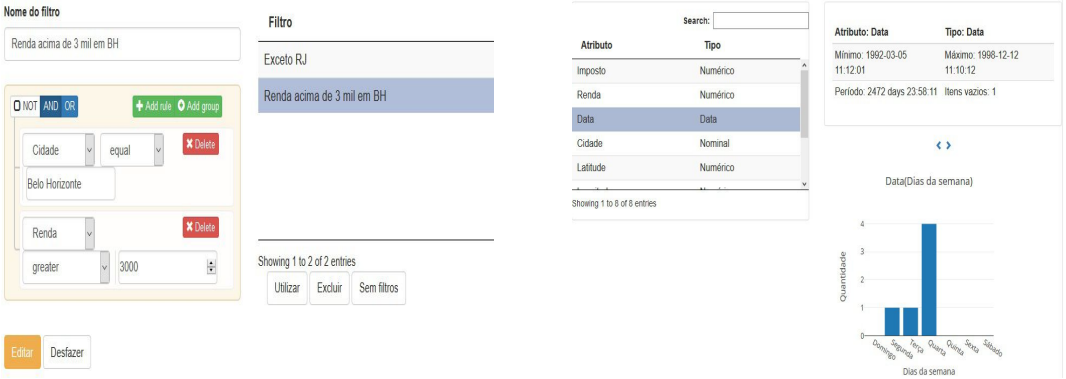
3.2. Georeferenciamento

Dados georreferenciados do sistema de coordenadas são compostos pela associação de atributos que correspondam a latitude e longitude. Dependendo dos dados utilizados, essa correspondência pode não estar representada explicitamente. Dessa forma, o usuário deve indicar quais atributos possuem a associação latitude/longitude, formando assim um atributo composto do tipo coordenada.

Uma vez definidos os tipos coordenadas, é possível visualizar os dados em um mapa (Veja Figura 2(b)). Algoritmos de aprendizado de máquina podem ser usados para realizar o agrupamento de dados do tipo coordenada. Nesse aspecto, a visualização das centróides encontradas através de pontos no mapa é essencial para que o usuário tenha um entendimento melhor sobre a sua base de dados.

3.3. Filtro

A ferramenta permite que o usuário filtre os dados, selecionando itens (linhas) com base nos valores dos atributos (veja Figura 3(a)). Os filtros são flexíveis, ou seja, podem ser editados após serem criados.



(a) Filtro (b) Estatísticas

Figura 3. Telas do DCluster

Um filtro é definido pelas associações de regras, que podem ser por meio das operações lógicas *NOT*, *AND* ou *OR*. Cada regra corresponde a um atributo associado

a uma operação e um valor, que pode ser numérico ou nominal, dependendo do tipo do atributo selecionado. As regras podem ser aninhadas, possibilitando que sejam elaboradas expressões lógicas complexas. O conjunto de operações para atributos numéricos são: diferente de, igual, maior, menor, maior ou igual, menor ou igual, intervalo de valores, e verificação de nulidade ou não nulidade de itens. Para atributos nominais, as operações são: igual, diferente, e verificação de nulidade ou não nulidade de itens.

3.4. Gráficos

O DCluster oferece um conjunto de gráficos para cada um dos tipos de atributos aceitos: numérico, nominal e data/hora (Veja Figura 3(b)). Para atributos numéricos, estão disponíveis os gráficos de barras e linhas. Os atributos data/hora e nominal possuem gráficos em barras, sendo que o último ainda permite o tipo de gráfico *pizza*. O gráficos têm a vantagem de serem interativos, permitindo que o usuário obtenha informações ao sobrepor o cursor sobre determinadas regiões. Além disso, os gráficos podem ser baixados, e caso o usuário necessite é possível realizar a edição no painel de edição de gráficos oferecidos pela biblioteca *Plotly*.

3.5. Estatísticas

Além dos gráficos, são geradas estatísticas descritivas conforme o tipo de cada atributo da base de dados. Para atributos numéricos são calculados a média, variância, mínimo e máximo. Para data/hora, são calculados os intervalos de valores, o mínimo e o máximo. No caso de atributos nominais, são geradas informações da quantidade de itens diferentes. Para todos os tipos de atributos, ainda é informada a quantidade de itens com o respectivo valor vazio ou nulo. Um exemplo é ilustrado na Figura 3(b).

3.6. Associações entre pares de atributos

Pares de atributos de diferentes tipos podem ser associados para a visualização de gráficos ou mapas. Os possíveis pares de atributos e seus respectivos gráficos são: Numérico x Numérico, gerando scatterplot; Nominal x Numérico, gerando gráficos de barra e boxplot; Data x Numérico, gerando gráficos de barra e boxplot; Coordenada x Numérico, gerando heatmap.

Tabela 2. Associações entre pares de atributos.

Eixo x / Eixo y	Numérico	Nominal	Data	Coordenada
Numérico	scatter	barplot e boxplot	barplot e boxplot	heatmap

3.7. Agrupamento

Para realizar o agrupamento de dados, o DCluster implementa uma versão paralela do algoritmo *K-means*. O conceito de paralelismo utilizado como base da implementação do algoritmo foi o de processos mestre/escravo, como apresentado no trabalho [Hadian and Shahrivari 2014]. O processo mestre inicialmente é responsável pela divisão dos dados em partes que são enviadas para os escravos processarem. Os processos escravos retornam as centróides encontradas para o mestre, que posteriormente utiliza

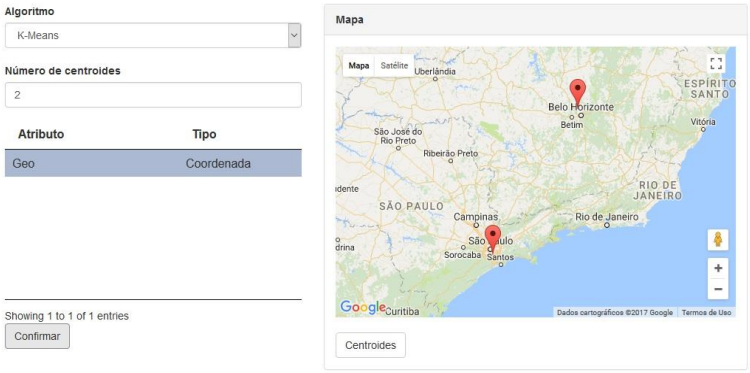


Figura 4. Agrupamento

esses dados recebidos como entrada do *K-means* para o processamento final. A Figura 4 apresenta um exemplo de agrupamento utilizando um atributo de coordenada.

Para avaliar a implementação paralela do agrupamento, foram realizados testes sobre duas bases de dados de diferentes tamanhos. Os resultados mostraram uma melhoria em desempenho de 26,6% para uma base de 73 MBytes e 35% para uma base de 310 MBytes, quando comparado com a versão tradicional do *k-means*.

3.8. Exportação

Após a exploração dos dados, o usuário poderá exportar os gráficos para gerar relatórios locais. O usuário também tem a opção de exportar a base de dados utilizada para o formato CSV. Essa é uma funcionalidade interessante quando se deseja obter os dados após a aplicação de um filtro.

4. Conclusão e Trabalhos Futuros

Este trabalho apresentou o DCluster, um sistema para a análise de grandes volumes de dados com foco em georreferenciamento. O processo de desenvolvimento do DCluster tem o objetivo de unir quatro importantes aspectos: suporte a dados georreferenciados, licença acessível, usabilidade e disponibilidade via Web. Esses fatores são essenciais para a popularização da análise de dados, afim de torná-la mais acessível a pesquisadores e analistas de dados que não possuem licenças de ferramentas pagas, mas que ao mesmo tempo precisam de um conjunto abrangente de funcionalidades.

Existem vários desafios a serem tratados no DCluster. Primeiramente, serão implementadas a customização no tratamento dos dados de entrada em diferentes formatos, criação de cubos de dados e suas dimensões, e na exportação de dados no formato *Planilha do Microsoft Excel - XLSX*. Posteriormente, será feita a integração do DCluster com ferramentas de Big Data como *Hadoop* e *Spark*. Por fim, serão disponibilizados outros algoritmos para análise de dados georreferenciados.

Com o objetivo de tornar o sistema financeiramente acessível, um outro desafio se refere à definição do modelo de negócios mais adequado para o DCluster. Nosso maior interesse é torná-lo acessível principalmente a estudantes e pesquisadores.

5. Agradecimento

Este trabalho teve o apoio financeiro da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Referências

BigML (2017). Bigml: Machine learning made easy. https://bigml.com/. Acessado em 01/07/2017.

Hadian, A. and Shahrivari, S. (2014). High performance parallel k-means clustering for disk-resident datasets on multi-core cpus. *The Journal of Supercomputing*, 69(2):845–863.

Hitachi (2017). Pentaho: Data integration, business analytics, and big data. http://www.pentaho.com/. Acessado em 01/07/2017.

Microsoft (2017a). Azure machine learning. https://azure.microsoft.com/pt-br/services/machine-learning/. Acessado em 01/07/2017.

Microsoft (2017b). Power bi: ferramentas do bi dde visualização de dados interativa. https://powerbi.microsoft.com/pt-br/. Acessado em 01/07/2017.

of Waikato, U. (2017). Weka 3: Data mining software in java. http://www.cs.waikato.ac.nz/ml/weka/index.html. Acessado em 01/07/2017.

Qlik (2017). Qlik: Business intelligence — ferramentas de visualização de dados. http://www.qlik.com/pt-br. Acessado em 01/07/2017.

SAS (2017). Sas: Software de business analytics e business intelligence. https://www.sas.com/en_us/home.html. Acessado em 01/07/2017.

Sisense (2017). Sisense: Business itenlligence (bi), software and analytics tools. https://www.sisense.com/. Acessado em 01/07/2017.

Statista (2017). Statista: the portal of statistics. https://www.statista.com/statistics/346195/facebook-global-mobile-dau/. Acessado em 01/07/2017.

Tableau (2017). Tableau: Análise e business intelligence. https://www.tableau.com/pt-br. Acessado em 01/07/2017.

Xavier, W. Z., Xavier, F. H. Z., and Marques-Neto, H. T. (2017). Visualizing and analyzing georeferenced workloads of mobile networks. In *IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 306–310.