Programa de Pós-Graduação em Engenharia de Sistemas

PROJETO DE DISSERTAÇÃO

Recife, Fevereiro de 2016.

Universidade de Pernambuco (UPE)
Escola Politécnica de Pernambuco (POLI)
Instituto de Ciências Biológicas (ICB)

MODELO PREDITIVO DE SUGESTÃO DE ROTEAMENTO RODOVIÁRIO DE CARGAS CONSIDERANDO DADOS HISTÓRICOS, FATORES SÓCIO-AMBIENTAIS E REDES SOCIAIS

Mestrando: Eng. Othon Luiz Teixeira de Oliveira
Orientador: Prof. Dr. Fernando Buarque de Lima Neto

Projeto de qualificação apresentado ao Programa de Pós-Graduação em Engenharia de Sistemas
Área de concentração: **Cibernética.**

Banca de qualificação:

Prof. Dr. Carmelo José A. Bastos Filho.....Engenharia de Sistemas/POLI/UPE

Prof. Dra. Rita de Cássia M. Nascimento.....Engenharia de Sistemas/POLI/UPE

Recife, Fevereiro de 2016.

"Quem escolheu a busca
não pode recusar a travessia"
(Guimarães Rosa)

Agradecimentos

Agradeço a fulano, ciclano e beltrano de tal,

Resumo

O Transporte de cargas, que atravessa as regiões metropolitanas das grandes cidades brasileiras o fazem principalmente pelas rodovias federais. Essas rodovias estão constantemente congestionadas e têm recebido aumento expressivo de novos veículos a cada ano.

No entorno desses trechos urbanos têm crescido desordenadamente comunidades que demandam por políticas sociais que atendam às suas necessidades. Para reivindicar, dos entes públicos, essas comunidades bloqueiam as rodovias, aumentando a pressão sobre os congestionamentos. Em alguns trechos o traçado das rodovias está próximo a morros e florestas ficando sucetíveis às intempéries climáticas. Essas variáveis impõem constantes paralisações às rodovias, representando atrasos nas entregas, custos adicionais às empresas e prejuízos à competitividade nacional.

Propor novas soluções, que venham minimizar esses constrangimentos é condição 'sine qua non', para mitigar o que vem sendo chamado de "Custo Brasil". Um sistema de suporte a decisão, ao roteamento de cargas, que utilize informações vindas de sítios eletrônicos como google-maps, base de dados históricas da Polícia Rodoviária Federal, das redes sociais e das condições socio-ambientais estão disponíveis em forma de dados na Internet e nos servidores locais dessas instituições, podendo ser processados numa plataforma auto-adaptável.

Assim, esse estudo tem por objetivo desenvolver um sistema de suporte a decisão para a problemática crescente da Logística de Cargas, antecipando prováveis eventos, que possam interferir no fluxo normal das rodovias brasileiras, propondo assim uma alternativa ao traçado de rotas determinísticas e incorporando novas opções de rotas inspiradas na predição de eventos futuros.

Palavras-chave: Mineração de dados, Bases de dados históricas, Redes sociais, Logística de transportes, Roteamento

Abstract

This project we propose to make a predict model recive information of the social networking, data bases, and highways managements, compiling everthing is avaliable , to check up routing a self-adapter plataform, in real-time, to drive a vehicle fleet for a destination, safely.

Keywords: Data Mining, Data Bases, Social Network, Logistic, Routing

Lista de Abreviações e Siglas

DM	*Data Mining*
BG	*Big Data*
PSO	*Particle Swarm Optimization*
ACO	*Ant Colony Optimization*
FSS	*Fish School Search*
B	*Byte*
KB	*kyloByte*
MB	*MegaByte*
GB	*GigaByte*
TB	*TeraByte*
PB	*PetaByte*
EB	*ExaByte*
ZB	*ZettaByte*
YB	*YottaByte*

Lista de Figuras

2.1 Planilha Survey 15
2.2 Técnica Map-Reduce 18
2.3 Big Data e Arquitetura Hadoop 19
2.4 Minerando dados no Big Data 20
2.5 Mapa mental da Mineração em textos 21
2.6 O padrão CRISP-DM 22
2.7 Domínio das técnicas aplicadas a mineração de dados 22
2.8 Minerando textos no Big Data com robôs 26
2.9 Árvore de decisão 27

3.1 Etapas da gerais da metodologia 29
3.2 Etapas 1 – Coleta e união das bases históricas de acidentes e interdições . . 32
3.3 Etapas da metodologia 33

Lista de Tabelas

2.1 Datas do Survey 16
2.2 Volume de dados no mundo 16
2.3 Matriz de Confusão 23
2.4 Matriz modelo de Confusão 23

3.1 Variáveis originais da base de acidentes 30
3.2 Variáveis originais da base de interdições 31
3.3 Variáveis do modelo preditivo 31
3.4 Cronograma – 12 meses 35

Sumário

1 Problema, Motivação e Objetivos **11**
1.1 Problema . . . 11
1.2 Motivação . . . 12
1.3 Objetivo Geral . . . 12
1.3.1 Objetivos Específicos . . . 12
1.4 Resultados Esperados . . . 13

2 Estado da Arte **14**
2.1 Introdução . . . 14
2.1.1 Detalhamento das etapas . . . 15
2.1.2 Map Reduce - Big Data . . . 18
2.1.3 Hadoop - Map-Reduce - Big Data . . . 19
2.2 Data Mining . . . 20
2.2.1 Data Mining - Big Data . . . 21
2.3 Enxame de partículas . . . 24
2.3.1 Data Mining - Swarm Intelligence . . . 24
2.3.2 Data Mining - Swarm Robotics . . . 26
2.4 Machine Learning . . . 26
2.4.1 Árvore de Decisão . . . 26
2.5 Regressão Logística . . . 27

3 Metodologia **28**
3.1 Plano geral da metodologia . . . 28
3.2 Proposição do modelo preditivo . . . 29
3.3 Reflexão sobre as tecnologias utilizadas no modelo preditivo . . . 29
3.3.1 Metodologia utilizada para coleta dos dados . . . 30
3.3.2 As variáveis do modelo preditivo . . . 31
3.4 Acoplamento com a estrutura dinâmica . . . 33
3.5 Cronograma . . . 35

4 Resumo da Proposta, Discussão **36**
4.1 Resumo da Proposta . . . 36
4.2 Discussão . . . 36

Referências Bibliográficas **38**

Referências Bibliográficas **38**

O Projeto

"E se o mundo não corresponde
em todos os aspectos a nossos desejos,
é culpa da ciência ou dos que querem
impor seus desejos ao mundo?"
(Carl Sagan)

1 Problema, Motivação e Objetivos

1.1 Problema

A partir do início do século XXI o mundo digital, especialmente a Internet, conheceu sua primeira grande crise (1). As empresas ligadas a esse mundo, conhecidas como PontoCom, para sobreviverem, adaptaram-se à Internet abrindo suas estruturas, desde então houve um *boom* nesse segmento. As informações geradas e disponibilizadas à Internet, nos mais recentes anos, representam 90% de tudo o que já foi criado nos anos anteriores pela humanidade ou desde que nossa civilização aprendeu a guardar informação.

Para armazená-los, seriam necessários milhões de computadores; se fosse possível dispô-los num único *DataCenter*, esses ocupariam uma área do tamanho do estado de São Paulo.

Os dados produzidos pelo ser humano atualmente dobram a cada 5 anos; esses dados são desde artigos publicados, novas técnicas para os mais diversos problemas da vida humana e outros, ficando impossível serem armazenados pelo cérebro humano (2).

Com a chegada da Internet das Coisas (IoC), acrônimo de *Internet of Things* (IoT), a previsão é de que o número de informações dobre a cada 2 anos.

A Internet da Coisas pode ser entendida como "coisas conectadas às coisas", que pode ser, o refrigerador comunicar-se com os alimentos ali depositados, que contenham uma etiqueta identificadora por rádiofrequência (Radio Frequency IDentification – RFID), podendo ter autonomia para enviar a um supermercado um rol de compras futuras.

Isso irá fazer com que os dados trafegados pela Internet tenham um crescimento exponencial. A essas informações circulantes dá-se o nome de *Big Data*. Um *Big data* é um conceito, na verdade são as coleções de tudo o que é disponibilizado na Internet, desde os dados dos *Data Centers* aos dados gerados pela Internet das Coisas. Essa enormidade de dados poderia ver a ser um problema de difícil solução se não houvessem abordagens computacionais com habilidade de extrair semântica bem como possibilidade de oferecer suporte a agentes decisores.

Uma instância do problema a ser tratado nessa pesquisa será a integração de bases heterogêneas de dados em um aplicação de suporte à decisão de logística de cargas rodoviárias.

1.2 Motivação

As rodovias federais que atravessam a Região Metropolitana do Recife (RMR) estão constantemente congestionadas, não apenas pela quantidade de veículos, mas por serem alvo de paralisações das mais diversas matizes, como protestos de trabalhadores, acidentes, buracos, intempéries naturais e outros tipos de paralisações. Em situações extremas poderiam paralisar até a produção das fábricas no seu entorno, por exemplo a Fiat - FCA (3). Esta será responsável por aproximadamente 1 000 caminhões cegonheiros nessas rodovias, quando do seu pico de produção (200 000 veículos/ano).

A RMR é a 5ª região mais populosa do Brasil, concentra 3.690.485 habitantes (dados de 2012) (4) em 14 municípios, além da RMR será considerada para a pesquisa a Zona da Mata Norte (ZMN) com 577.191 habitantes e a Zona da Mata Sul (ZMS) com 733.447 habitantes. Nessas regiões (RMR, ZMN e ZMS) a frota (automóveis particulares, ônibus, caminhões, motocicletas, tratores e outros veículos) foi contabilizada, em 2014, com 635.686 veículos (5).

O que acontece na região metropolitana do Recife, é frequentemente visto no entorno das principais cidades brasileiras. Por outro lado, câmeras de monitoramento de trânsito, redes sociais, aplicativos de celular e outros dispositivos, fornecem informações diárias sobre o que acontece nessas rodovias e no entorno delas, atualizando e alimentando bases de dados históricas, em repositórios espalhados pelos centros de monitoramento de trânsito, isso é o *Big data*.

A proposição de uma solução para resolver esse problemática requer várias etapas, para além da proposição de algumas técnicas de mineração dos dados. Propomos, nesse projeto, uma solução peculiar, ao enviar essa frota de caminhões por diversas rotas, escolhidas por critérios cientificamente estudados. Isso poderá ser de suma importância para solucionar a problemática do transporte de cargas, que poderá advir com a plena produção da FCA, e que não se aplica apenas à região metropolitana do Recife, mas a toda e qualquer fábrica do país. Permitirá fornecer toda informação que se faz necessária para acompanhar os caminhões na transposição dos obstáculos que possam surgir ao transitar por Pernambuco, conduzindo-os até seu destino de maneira segura e no menor tempo possível.

1.3 Objetivo Geral

Esse projeto de pesquisa tem como objetivo desenvolver um Modelo de sistema de Suporte à Decisão para a problemática das retenções crescentes de logística de cargas rodoviárias. Para isto propomos uma solução multidisciplinar através da integração de diversas tecnologias disponíveis desde a análise dos dados históricos das rodovias até a situção cotidiana.

1.3.1 Objetivos Específicos

- Representar a problemática da logística de cargas em uma plataforma adaptável:
    1. Desenvolver uma plataforma adaptável que utilize dados de redes sociais (Twitter); analisar o contexto das rodovias por essas redes através da mineração de dados em textos e integrar ao resto do sistema através que verifica palavras chaves como: protestos, acidentes e outras.
- Desenvolver um modelo preditivo dos fenômenos que envolvem as rodovias:

1. Para o desenvolvimento do modelo preditivo pretende-se utilizar bases de dados históricas.
2. O modelo preditivo integra várias bases de dados, tais como: Polícia Rodoviária Federal – PRF, Batalhão de Polícia de Transito – BPRv e dados históricos do Instituto de Pesquisas Espaciais – INPE ou dos dados de precipitações pluviométricas do "European Centre for Medium-Range Weather – ECMWF.
3. Algumas dessas informações também estão disponíveis em base de dados abertas, como sugere o Portal da Transparência, nos servidores da PRF além de outras informações para complementar o sistema estão disponíveis na Internet sendo atualizadas pela PRF através de uma API aberta, esta pode ser configurável para se ligar ao nosso sistema.

- Propor uma simulação interativa da estrutura com a dinâmica:
    1. A malha viária está representada na Internet em mapas de bases vetoriais que pretendemos integrar ao nosso sistema de informações e também vamos incorporar as informações que a Polícia Rodoviária Federal(PRF) dispõe, que controla as rodovias BRs.
    2. A simulação interativa utiliza uma plataforma baseada na API do Google Maps.
    3. Para simulação interativa da estrutura preditiva com a dinâmica real serão capturados "feeds" de redes sociais, por exemplo pelo Twitter. Essa técnica fará um arco cibernético mantendo o sistema preditivo atualizado.

1.4 Resultados Esperados

Ao final dessa pesquisa pretende-se obter um Modelo de Sistema de Suporte à Decisão adaptável, que contribua como uma ferramenta para complementar à tomada de decisão para a gestão do transporte de cargas rodoviárias. A região inicialmente escolhida será a RMR.

2
Estado da Arte

2.1 Introdução

Astrônomos atualizam suas descobertas numa base de dados disponíveis para outros utilizarem; as ciências biológicas agora têm tradição em depositar seus avanços científicos em repositórios públicos; redes sociais estão focadas na Web: Facebook, LinkedIn, Tweeter e outras sobrevivem coletando informações, vendendo espaço publicitário e repassando-as às empresas de telemarketing, empresas de comércio eletrônico como Amazon.com, Submarino.com.br, Americanas.com, MagazineLuíza.com.br, utilizam essas informações para vender mais e melhor. Por outro lado, artigos científicos dos mais variados assuntos e das mais variadas áreas alimentam todos os dias, com milhões de informações, os *Data Centers*. Esse é o paradigma do *Big Data*.

Nesse estudo, para fazer frente ao paradigma acima aludido e extrair informações com eficácia, foi proposto o desenvolvimento de um mapeamento sistemático.

Mapeamento sistemático

Foi desenvolvido a priori um mapeamento sistemático das mais diversas tecnologias empregadas no "Big Data" e seus paradigmas para extrair informações e conhecimentos, analisando as técnicas de Inteligência Artificial mais empregadas nesse campo científico de exploração. A posteriori; propor uma solução à logística de cargas aplicada à realidade brasileira, antecipando os problemas que possam surgir no traçado de rotas determinísticas.

O mapeamento sistemático foi definido em etapas da seguinte forma:

A. Coleta dos dados;
B. Seleção dos artigos;
C. Escolha dos Filtros;
D. Leitura dos artigos.

2.1.1 Detalhamento das etapas

A. Coleta dos dados

Para coletar os dados foram escolhidas seis palavras-chave com relevância para o tema, a saber:

- “Data Mining and Swarm Intelligence”
- “Data Mining Big Data”
- “Data Mining Swarm Robotics”
- “Hadoop Map Reduce in Big Data”
- "Machine Learning"
- “Map Reduce Big Data”

A partir das palavras-chave foram criadas planilhas, sendo cada aba representada por uma palavra-chave. Incluiu-se no mínimo 30 artigos para cada aba da planilha, totalizando 316 artigos. Pretendeu-se, com essa técnica, construir rapidamente gráficos das mais diferentes matizes, tais como:

- Base pesquisada;
- Data da publicação;
- Parecer (Aceito ou rejeitado);
- Título do artigo;
- País de origem.

A ferramenta utilizada para obter os dados referentes às palavras-chave (referência bibliográfica) foi o programa Mendeley, especializado em extrair dados de arquivos como artigos, dissertações e teses. A figura 2.1 a seguir mostra como a Planilha-Survey se apresenta:

Figura 2.1: Planilha Survey

	A	B	C	D	E	F
1	Base Pesquisada	ata da Publicaç	ata pesquisad	(A)ceito / (R)ejeitado	Título	Pais
2	Journal of Advanced Co	3/5/2014	19/5/2015		Swarm Algorithm for Unmanned /	Filipinas
3	Malware Detection	3/1/2007	19/5/2015		An Inside Look at Botnets	Estados Unidos
4	IEEE International Sym	3/1/2011	19/5/2015		Accelerating the Nussinov RNA f	Estados Unidos
5	Robotics and Computer	10/9/2001	19/5/2015		Rapid response manufacturing th	Singapura
6	4th International Confere	3/1/2009	22/5/2015		Botnet: Survey and case study	China
7	Conference Proceeding:	3/1/2008	22/4/2015		Applications and prototype for sy	Estados Unidos

B. Seleção dos artigos

A seleção dos artigos foi primeiramente escolhida pelas mais recentes publicações dos últimos 5 anos, expecificamente entre 2010 e 2015. O segundo critério de seleção foi pelo órgão publicizador, considerando os mais conhecidos, tais como IEEE, Elsevier, Springer, bem como os jornais a eles pertencentes. Outras fontes foram tabém consideradas, como universidades e seminários mais conhecidos na área. Outro critério considerado foi a seleção de no mínimo de 30 (trinta artigos) por palavra-chave. Dessa forma, procuramos aproximar os dados coletados da distribuição normal padrão, que por razões evidentes está amplamente tabelada e é suprida por quase todos os softwares para construtores de gráficos.

C. Critérios de Inclusão/Exclusão
O critério de inclusão e exclusão foi baseado na leitura dos "abstracts" dos artigos. Devido ao fato de algumas palavras-chave estarem muito em evidência, as mesmas são citadas em diversos artigos sem referência para tema, não sendo relevante para a pesquisa. Dessa forma foi criada uma pasta chamada "Rejeitados" para onde foram movidos esses artigos. Outro critério de exclusão foi data anterior a 2010, com exceção dos artigos clássicos da área.

D. Leitura dos artigos
A leitura dos artigos iniciou-se após a etapa de Inclusão/Exclusão. Até o momemto foram lidos 72 artigos de um total de aproximadamente 240 artigos. Segue uma tabela com as datas do plano de execução.

Tabela 2.1: Datas do Survey

Data/2015	**(A) Coleta**	(B) **Seleção**	**(C) Filtros**
Abril a Junho	x	—	—
Junho e Julho	–	x	x
Agosto	–	–	x

Big data

"Em 2010 empresas e usuários armazenaram mais de 13 exabytes de novos dados"(6). O *Big Data* pode ser definido como um paradigma dos 3 V's, descrito como:

- Volume de dados;
- Velocidade para acessar esses dados;
- Variedade de informações;

Dentre as mais diversas técnicas encontradas para acessar dados no *Big Data* destacamos o Map Reduce, da "framework" Hadoop, por ser de livre uso (freeware) e por ser tolerante a falhas, sendo de fácil implementação e de baixo custo para a maioria das pequenas empresas e pesquisadores. A tecnologia congênere, desenvolvida pelo Google, ("Google File System") merece ser destacada. Não é tarefa trivial inferir e desenvolver novas ferramentas para o *Big Data*, dada a dimensionalidade do problema, conforme destacado na tabela a seguir:

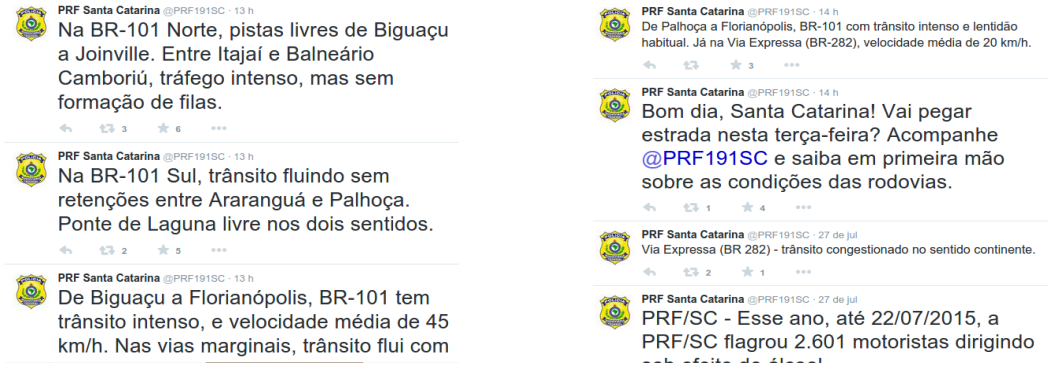
Tabela 2.2: Volume de dados no mundo

Ano	**Qtd**	**Unidade**	**Múltiplo**
2000	800	terabytes – TB	10^{12}
2006	160	petabytes – PB	10^{15}
2009	500	exabytes – EB	10^{18}
2012	2,7	zettabytes – ZB	10^{21}
2020	35	yottabytes – YB	10^{24}

Com a chegada da Internet das Coisas, conhecida como “A terceira onda da Internet”, os dados irão aumentar exponencialmente. Por exemplo, os produtos nas gôndolas do supermercado, com um novo tipo de etiqueta se conectam a uma leitora de radio frequência a alguns metros de distância, contabilizando o total do estoque em segundos; o consumidor leva seus produtos escolhidos ao caixa desse supermercado, pagando a conta sem precisar retirar qualquer produto do carrinho. Ao introduzir esses produtos na geladeira, será possível saber quando expirará a data de validade de determinados produtos, sem precisar abri-lo, e quando acabarem esses produtos, a própria geladeira informará ao supermercado a falta deles, reservando o próximo rol de compras. Assim poderá se dar essa onda de coisas conectadas, que fará com que os dados sofram emplosão combinatória de informações, multiplicando exponencialmente as dimensões do *Big Data*.

As redes sociais são um arcabouço de informações sobre todo tipo de assunto vivenciado no cotidiano das pessoas, inclusive situações que dizem respeito ao nosso ambiente de pesquisa. O cenário abaixo, encontrado numa rede social, exemplifica a sequência de informações retiradas do Twitter, um microblog onde os usuários escrevem num pequeno espaço (cerca de 140 caracteres), os mais diversos assuntos, e os usuários conectam por uma multiplicidade de dispositivos: computadores, tablets e celulares, formando uma grande rede social mundial. A ideia inicial do Twitter, segundo seus fundadores, era que essa rede se comportasse como um “SMS da Internet” (7). As informações são enviadas aos usuários, conhecidas como twittes, em tempo real e também enviadas aos usuários seguidores que tenham assinado para recebê-las.

A seguir pode-se verificar uma sequência de twittes da Polícia Rodoviária Federal de Santa Catarina:



A Polícia Rodoviária Federal de Santa Catarina, disponibilizou às 13h através do canal @PRF191SC, informações relevantes sobre o trânsito naquela localidade, num espaço temporal variado, por exemplo: entre Itajaí e Balneário Camboriú o transito está intenso. Isso sugere que a frota de caminhões deva ter uma rota alternativa caso a situação persista por muito tempo. No primeiro twitte da segunda coluna, é informado em Via Expressa (BR 282) que o trânsito está lento com velocidade de 20km/h (praticamente congestionado). Essa informação sugere que deve ser pensada uma rota alternativa, caso o congestionamento persista por muito tempo.

Outra rede social conhecida pelos condutores de veículos é o Waze. O Waze é um aplicativo de navegação para o trânsito, funciona em aparelhos celulares e tablets. Os utilizadores desse aplicativo são conhecidos como wazers e compartilham informações sobre o trânsito, em tempo real. Toda via, as informações somente estão disponíveis no momento em que são postadas pelos utilizadores, por um período de tempo pequeno. Caso não haja usuários trafegando pelas vias ou caso os mesmos não tenham disponibilidade

em postar informações, não há o que se compartilhar. Outro problema levantado com o waze é que, caso não haja conexão à Internet não há como acessar os dados dos 'wazers', para navegação.

Além dos dados que chegam ao *Big Data* através das redes sociais, as grandes cidades têm disponíveis câmeras de monitoramento do trânsito nos semáforos ou próximas a eles; algumas com cobertura por canais de televisão bem como câmaras de segurança próximos às rodovias, coletando informações em tempo real. Os dados desses dispositivos são gravados, sendo conhecidos como *stream* de dados. Esses *streams* podem ser disponibilizados na Internet, em sítios eletrônicos especialmente construídos para isso, como o http://vejoaovivo.com.br dentre outros.

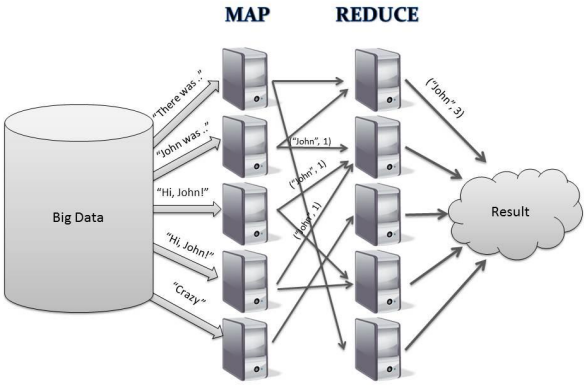
Os dados disponibilizados pelos diversos meios de comunicação não estão em formato que possam ser utilizados imediatamente, precisando antes serem processados. Tais dados não processados são conhecidos como "dados frios". O processo de tratar as informações, retirando-lhes o "lixo" e transformando dados "frios" em dados "quentes", é um processo que tem um custo temporal elevado, devido ao volume dos dados.

Para trabalhar com os dados do *Big Data* a empresa Google (Google Inc.) criou uma arquitetura para computadores trabalharem remotamente em conjunto, formando um *cluster*. Essa arquitetura reestrutura o sistema de arquivos, o *filesystem* [1] (8) e os diretórios dos computadores com certas características, conhecido como HaDoop FileSystem (HDFS). O Hadoop é uma implementação em código aberto da técnica Map-Reduce.

2.1.2 Map Reduce - Big Data

O Map-Reduce é uma técnica conhecida desde a linguagem Lisp, implémentado sob uma arquitetura HDFS (ou GFS) que permite mapear os *clusters* e reduzir a quantidade de informação conforme um par de informações <chave:valor>. A imagem a seguir exemplifica essa técnica:

Figura 2.2: Técnica Map-Reduce



Map-Reduce; modelo de computação paralela para ser utilizado na Internet foi proposto pelo Google (9).

[1] *Filesystem* referem-se à forma como os dados são armazenados, organizados e acessados, pelo sistema operacional, em cada partição no disco (ou no disco rígido inteiro)

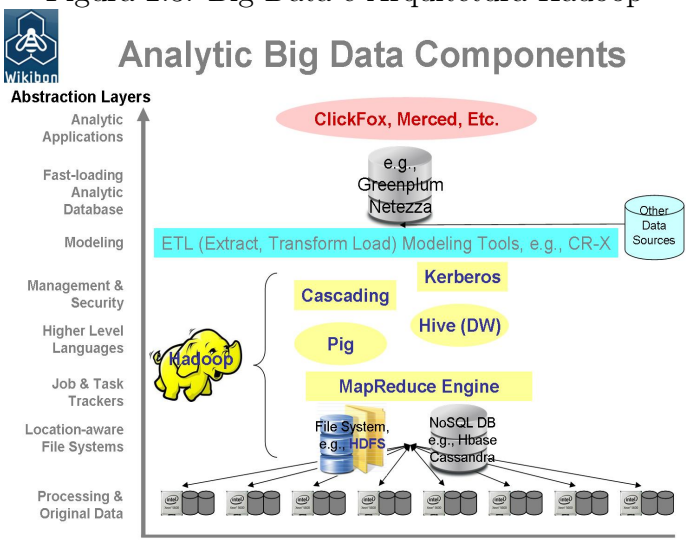
2.1.3 Hadoop - Map-Reduce - Big Data

O Hadoop é uma arquitetura de milhares de computadores interligados paralelamente e espalhados pela Intenet formando *cluster*. Esse *cluster* tem a característica de grande escalabilidade, em torno de 3 000 computadores, dependendo da construção e tolerância a falhas. Quando um computador do *cluster* fica inoperante ou "cai", os dados são salvos em outro computador.

A estrutura de diretórios do Hadoop foi especialmente construída para lidar com as características descritas anteriormente, devido ao baixo custo de implementação e por ser uma tecnologia aberta (*Open Source*), tendo sido desenvolvida por programadores do mundo todo. A versão Hadoop da empresa Google é o Google Filesystem (GFS).

Os *clusters* de computadores, aplicando a técnica de programação Map-Reduce estão preparados para extrair dados do *Big Data*, contudo até que esses dados estejam prontos para mineração eles devem seguir um fluxo de operações para transformá-los em dados relevantes (quentes). O fluxo seguido pela informação, desde o local onde é produzida até o momento em que possa ser utilizada está dividido em etapas na imagem a seguir:

Figura 2.3: Big Data e Arquitetura Hadoop



Fonte: Wikibon – 2015

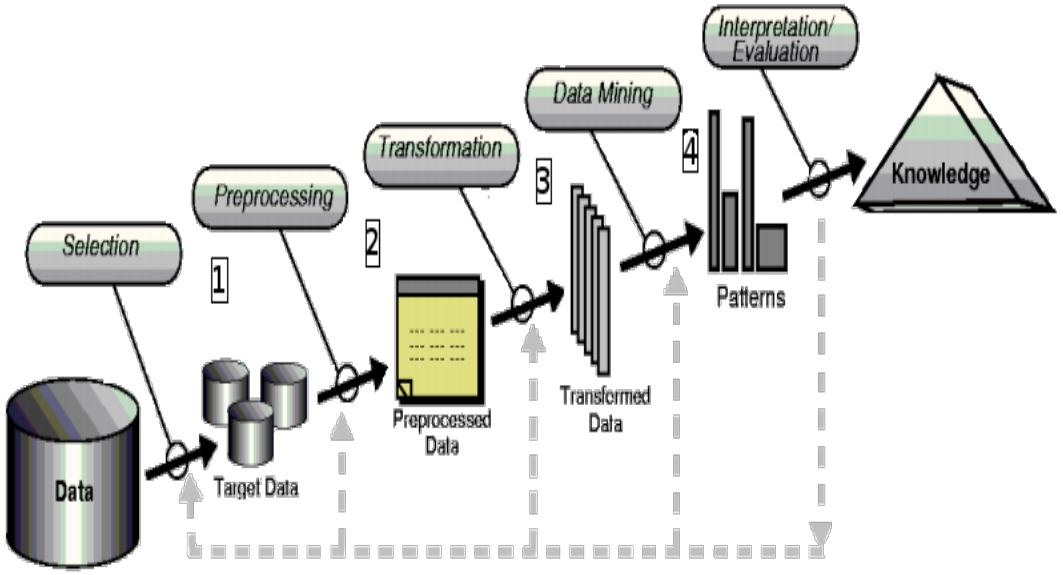
A camada mais baixa da imagem, onde se lê *Processing & Original Data* é a origem do *Big Data*, onde estão os dados "frios". Entre as camadas *Location-aware File Systems* e *Management & Security* a informação é mapeada no Hadoop e em seguida a técnica Map-Reduce é empregda, reduzindo o volume dos dados. Em Hive (Datawarehouse - DW) é possível armazenar os dados, esses dados são considerados "quentes". Em *Modeling* ocorre o que se chama de *Extract, Transform and Load* (ETL) ou Extração, Transformação e Carga, que é onde se inicia a mineração de dados. Mais adiante aprofundaremos a discussão sobre Mineração de dados.

2.2 Data Mining

Técnicas de mineração de dados trabalham com dados estruturados, preenchidos em sua totalidade (sem *missing data*), para poder extrair informações relevantes. Um dos maiores problemas na extração de informações é que os dados não estão estruturados, ou não estão no grão adequado, ou ainda faltam dados ("missing data" – dados ausentes). Para contornar o problema de dados ausentes existem várias técnicas, como preenchimento dos dados através de técnicas de inteligência artificial.

O caminho da extração dos dados até sua mineração e, por fim, extração de conhecimento é longa. Na figura a seguir temos um exemplo desse caminho:

Figura 2.4: Minerando dados no Big Data



(Excerto de Fayyad *et al.*, - 1996)

O Big data está representado na figura onde se lê "Data" e está repleto de *missing data* e/ou dados inconsistentes, conhecidos como dados não estruturados. O balão onde se lê "Selection" representa a coleta das informações ou a seleção dos dados no *Big Data*. Em nossa pesquisa esses dados são provenientes das mais diversas fontes, tais como, redes sociais, câmaras de trânsito, informações de satélites meteorológicos e outras fontes.

Armazenar qualquer quantidade de dados nessa etapa pode ser um grande problema, devido à sua extensão, porém os dados relevantes podem ser armazenados em "Target Data" com a tecnologia Hadoop, utilizando as técnicas de "Map" e "Reduce" para criar *cluster* de informações e ler os fluxos de dados (stream data). Algumas técnicas de IA podem ser aplicadas nessa etapa como, "Data Mininng Swarm Robotics" através de Botnets [2] ou "Swarm Intelligence".

No balão "Preprocessing" os dados não-estruturados são tratados, por exemplo, retirando os *missing data*. Para estruturar as informações é preciso utilizar técnicas linguísticas, uma vez que existe lógica entre eles (10). Esses dados normalmente são coletados por técnicas de Mineração de Textos, também conhecidas como Mineração de Dados em

[2]Botnet é citado no sentido da coleta de informações

Textos, técnicas de IA como "Machine Learning" têm sido muito utlizadas. Em "Transformation" os dados foram em estruturados, podendo ser armazenados em Bancos de Dados, conhecidos como Datawarehouse, por exemplo o Hive.

O processo de Mineração dos dados começa no balão "Data Mining", onde são aplicadas as técnicas de IA conhecidas como classificadores, para extração de padrões, tais como: "Decision Tree" (Árvore de decisão), "Artificial Neural Network" (Redes neurais artificiais), "Logistic Regression" (Regressão Logística) e "Deep Learning". Algumas técnicas de mineração de dados são fortemente influenciadas pelas informações na entrada (input), como as Árvores de decisão (11). As Redes Neurais, dependendo da quantidade de variáveis de entrada, paderão ter milhares de neurônios na camada intermediária, o que inviabilizaria essa metaheurística [3].

Todas essas etapas descritas na figura são recorrentes, como indicam as setas pontilhadas que retornam aos passos anteriores. Utilizar técnicas de mineração de dados, além de extrair dados, extrai conhecimento, com isso pode-se predizer os resultados futuros na saída do modelo, quando determinados dados ocorrem na entrada (12), essa técnica de extração de conhecimento chama-se *Knowledge Discovery Databases* (KDD).

2.2.1 Data Mining - Big Data

Minerar dados no *Big data* não é uma tarefa atômica, devendo ser divida em várias etapas, com processos específicos em cada uma delas, como descrito anteriormente. Extrair conhecimento dos dados não processados não faz sentido, tratá-los apenas "per si" exige muito trabalho de IA, como Mineração de dados em textos. A Mineração em textos é inspirada em técnicas de "Machine Learning" (10). Contudo analisar textos é basicamente entender o significado do texto, baseado em regras de associação lógica. O mapa mental a seguir mostra um modelo de análise de texto feito por seres humanos.

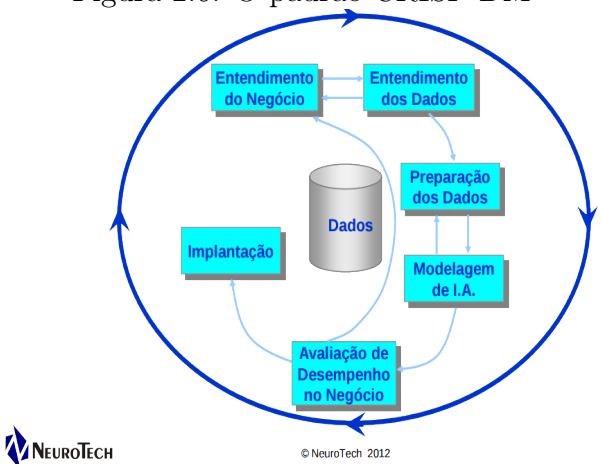
Figura 2.5: Mapa mental da Mineração em textos

O processo de mineração "CRoss Indrustry Standard Process for Data Mining" (CRISP-DM) (13) descreve como os especialistas em mineração de dados aplicam as técnicas em lide. O CRISP-DM é um processo recursivo, onde cada etapa deve ser revista até quando

[3]Metaheurística são heurísticas aplicadas em problemas onde os custos computacionais não são tratáveis em tempo polionomial, devido às explosões combinatórias geradas pelo grande número de tentativas. Metaheurísticas bioinspiradas metaforizam o comportamento de animais sociais, tais como formigas, pássaros, peixes e outros

o modelo apresentar os resultados satisfatórios, preliminarmente definidos. O Analista de Dados ou o Cientista de Dados é o profissional que acompanha e executa o modelo de predição, como exemplificado na figura a seguir:

Figura 2.6: O padrão CRISP-DM

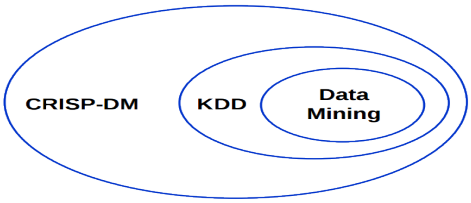


O **Entendimento do negócio** é uma fase crucial da mineração, um especialista (ou muitos) deve ser consultado. O analista de dados consegue fazer re-uso de conhecimento lendo periódicos e artigos, mas a experiência de um profissional da área é condição "sine qua non" nessa fase. Após a primeira fase, o analista de dados à fase do **Entendimento dos dados**. Nessa fase o analista "olha" para os dados com a acurácia de um especialista, procurando identificar qualidade nos dados. Dados ausentes são comuns em bases de dados não estruturadas, os "missing data" são sempre um problema a ser considerado, pois seu tratamento pode consumir muito tempo do analista de dados. A fase da **Preparação dos dados**, cobre a construção final do conjunto de dados. Preparar os dados significa criar e selecionar atributos, criar tabelas ou planilhas e registros dos dados.

Na fase de **Modelagem de I.A.** a tecnologia deve ser escolhida com critério baseado em experiência do analista de dados. Em sistemas de suporte à decisão, uma tecnologia inadequada pode levar a decisões imprecisas. É comum retornar às fases anteriores para adequar a técnica aos dados. Um modelo de regressão logística para problemas binários, redes neurais para problemas de classificação e assim por diante.

Na fase da **Avaliação de desempenho** um ou muitos modelos devem ter sido construídos e testados, apresentando alta qualidade da perspectiva da ánalise dos dados. Criar um modelo geralmente não é o fim do processo, contudo é um modelo de entendimento do negócio. Até que se obtenha as respostas satisfatórias, o modelo deverá ser refeito várias vezes até sua **Implantação**. A próxima figura descreve o domínio das técnicas aplicadas à mineração de dados:

Figura 2.7: Domínio das técnicas aplicadas a mineração de dados



Fonte: Neurotech – 2012

Quando são desenvolvidos sistemas de predição e ánalise de diagnóstico, há que se avaliar o desempenho e a qualidade dos resultados encontrados. Um método gráfico eficiente para detecção e avaliação da qualidade de sinais, conhecido como *Receiver Operating Characteristic* – ROC, ou curva ROC (14), foi criado e desenvolvido na década de 50 do século passado, para avaliar a qualidade da transmissão de sinais em um canal com ruído. Recentemente a curva ROC tem sido adotada em Mineração de dados e Aprendizagem de Máquina (15), em sistemas de suporte à decisão na medicina, para analisar a qualidade da detecção de um determinado teste bioquímico, na psicologia para detecção de estímulos (16) em pacientes, e na radiologia para classificação de imagens.

Essas métricas são amplamente utilizadas na classificação binária de resultados contínuos. Para isso ser construido, a Matriz de Contingência classifica as probabilidades como: verdadeiro positivo, falso positivo, falso negativo e verdadeiro negativo, respectivamente *True Positive – TP, False Positive – FP, False Negative – FN e True Negative – TN* , também conhecida como matriz de confusão, descrita na tabela a seguir:

Tabela 2.3: Matriz de Confusão

	Predito	
Real	TP FN	Positive – POS
Real	FP TN	Negative – NEG
—	PP PN	—

Tabela 2.4: Matriz modelo de Confusão

	Y Y	
X	P(X,Y) P(X,$\bar{Y}$)	Positive – POS
$\bar{X}$	P($\bar{X}$,Y) P($\bar{X}$,$\bar{Y}$)	Negative – NEG
—	P(Y) P($\bar{Y}$)	—

A matriz da Tabela 2.3 sintetiza a matriz da Tabela 2.4, portanto as duas tabelas são equivalentes. De acordo com as probabilidades condicionais temos:

$$P(X,Y) = P(X|Y).P(Y) = P(Y|X).P(X) \tag{2.1}$$

Então, a taxa de verdadeiros positivos será $P(Y|X)$ e a probabilidade de falsos alarmes ou taxa de falsos positivos será $P(Y,\bar{Y})$, a barra sobrescrita em $\bar{X}$ (ou $\bar{Y}$) representa negação.
A curva ROC será construída cruzando-se a taxa dos verdadeiros positivos (tpr = P(Y|X)) com a taxa dos falsos positivos (fpr = P(Y,$\bar{X}$)).

2.3 Enxame de partículas

Em 1989, G. Beni e J. Wang cunharam a expressão *Swarm Intelligence*, no seu trabalho em Robotic Swarm (17). O estudo do reino animal aprofundou-se na análise comportamental e possibilitou o melhor entendimento de como cooperam indivíduos dentro de um grupo, e quais os mecanismos usados para controlar o enxame, bem como, condicionar o indivíduo, tais como a estigmergia [4]. Por enxame, pode se entender manada, alcateia, bando, colônia, entre outras designações conforme o animal ou inseto e, a partir daqui, qualquer referência a um grupo de agentes passa a ser feita por enxame, por exemplo, um enxame de pássaros. Os cinco princípios da inteligência de enxame, segundo Chambers (18), são:

- Proximidade: os agentes têm que ser capazes de interagir;
- Qualidade: os agentes devem ser capazes de avaliar seus comportamentos;
- Diversidade: permite ao sistema reagir a situações inesperadas;
- Estabilidade: nem todas as variações ambientais devem afetar o comportamento de um agente;
- Adaptabilidade: capacidade de se adequar às variações ambientais.

A optimização por colônia de formigas ou *Ant Colony Optimization* - (ACO) é uma técnica de otimização por enxames que foi introduzida desde os anos 90's (19) baseado no comportamento forrageiro de colônias de formigas. O comportamento forrageiro em diversas espécies (20) é objeto de estudo das ciências biológicas, pois os animais predadores procuram otimizar seu ganho de proteína ao comer sua presa, minimizando o gasto de energia, ou minimizando o esforço para caçar, capturar e comer a presa. Esse comportamento é explorado pelo ACO para buscar soluções aproximadas para um problema de otimização discreto, para problemas de otimização contínuos e para problemas de roteamento em telecomunicações.

No caminho da busca por alimentos, as formigas deixam no ambiente uma marca chamada feromônio. Esse feromônio evapora com o passar do tempo, sendo assim, à medida que mais formigas sigam determinado caminho, mais intenso o feromônio se fará presente.

A equação da evaporação do feromônio pode ser:

$$p(i,j) = \frac{[\tau(i,j)]^{\alpha}.[\eta(i,j)]^{\beta}}{\sum[\tau(i,j)]^{\alpha}.[\eta(i,j)]^{\beta}} \tag{2.2}$$

2.3.1 Data Mining - Swarm Intelligence

Recentemente algorítimos de classificação baseados em "Ant Colony Optimization" têm sido experimentados em mineração de dados(21), o AntMiner é um deles.

O algorítimo AntMiner utiliza as formigas para gerar regras de classificação. Inicia com regra vazia e incrementalmente adiciona regras, uma de cada vez. A adição de cada termo é probabilística e baseada em dois fatores: a qualidade heurística do termo e a quantidade de feromônio depositado anteriormente pelas formigas. Após a parte antecedente da regra ser construída, a parte consequente da regra é assinalada por maior votação da amostra de treinamento coberta pela regra. A regra é construída com podas

[4]Estigmergia é a capacidade de insetos sociais trabalharem organizados sem a necessidade de planejamento nem controle central

aos termos irrelevantes, para melhorar a acurácia do algorítimo. A qualidade da regra construída é determinada e o valor do feromônio é atualizado na trilha pela formiga, proporcional à qualidade da regra. Quando todas as formigas construírem as suas regras, as melhores regras serão selecionadas, colocadas numa lista de regras descobertas. A amostra de treinamento corretamente classificada pela regra é removidas do conjunto de treinamento. Esse processo é continuado até que o número de amostras não cobertas pela regra seja pequeno e o usuário estabeleça um limiar. O resultado é uma lista de regras ordenadas que será usada para classificar um conjunto de testes. A seguir apresentamos o algoritmo de classificação AntMiner:

Algoritmo 1: ANT-MINER

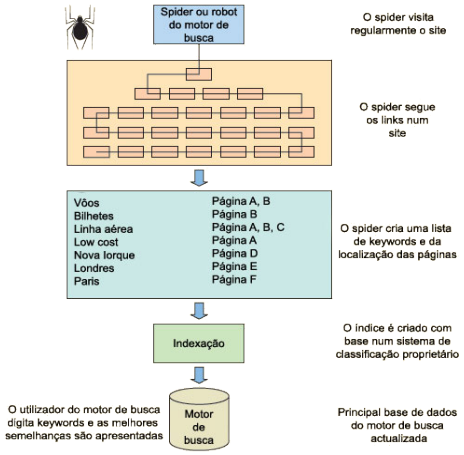
Entrada: $conjTreino = \{todosCasosTreinamento\}$;
Saída: $DescobrirListaRegras = []$;

```
1  início
2      enquanto conjTreino > MaxNaoDescoberto faça
3          t = 1 /* índice de formmigas */
4          j = 1 /*índice de convergência */
5          /*inicializa todas as trilhas com algum feromônio*/
6          repita
7              Ant_t /*inicializa com regras em branco e incrementalmente constrói um
                classificados de regras R_t por adição de termos um a um para as regras
                correntes; */
8              Podar regras R_t; Atualizar o feromônio de todas as trilhas
                incrementando o feromônio na trilha de cada Ant_t (proporcional à
                qualidade do R_t) e decrementa o feromônio em outras trilhas
                (simulando a evaporação de feromônio) se R_t == R_t − 1 /* atualiza a
                convergência de testes */
9              então
10                 j = j + 1;
11             senão
12                 j = 1;
13             fim
14         fim
15         t = t + 1;
16     até t >= No_ants ou j >= NoRegrasConverg;
17     Escolher melhor regra R_bost entre todas as regras R_t construída pelas
        formigas; Adicione regras R_bost para DescobrirListaRegras; conjTreino =
        conjTreino - {casos correntes cobertos por R_bost};
18 fim
19 fim
```

2.3.2 Data Mining - Swarm Robotics

O rápido crescimento da Internet tem trazido a reboque o contínuo crescimento da insegurança nos computadores e sistemas atuais (22). Ataques utilizando computadores através de robôs, conhecidos como Botnet, é um problema constante para quem lida com segurança da informação. No entanto os Botnets podem ter uma vida mais digna, como coletar informações no Big data. O Google se especializou nisso quando utiliza seus "Spiders". Esses robôs navegam pela Internet "devorando" páginas e indexando-as para que as buscas do motor de buscas do Google seja mais eficiente. Na figura a seguir podemos ver um exemplo desses robôs: [5] (23)

Figura 2.8: Minerando textos no Big Data com robôs



A utilização desses robôs é uma alternativa eficiente para coletar informações no Big data uma vez que essa é sua especialidade. Como o Big data tem uma forte componente de inconsistência, coletar informações das páginas visitas pelos robôs ou mesmo coletar a página inteira, pode ser extremamente eficiente nos quesitos "volume" e "velocidade" descrito na seção 2.3.2 (Hadoop – Map Reduce – Big Data).

2.4 Machine Learning

O desafio dos 3 V's (Velocidade, Variedade e Volume) pode ser um fator determinístico na escolha da ferramenta mais adequada para analisar dados do *Big Data* e extrair informação. As árvores de decisão são algoritmos rápidos, contudo dados impuros podem comprometer o desempenho desse algoritmo. A fase de extração dos dados do *Big Data* é fortemente influenciáveis pelas variáveis escolhidas, (11) isso pode representar o desafio maior para implementar esta técnica.

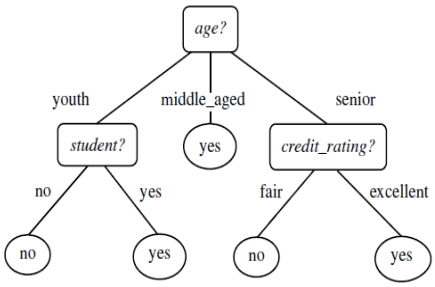
2.4.1 Árvore de Decisão

Han e Kamber (24) definem indução por árvore de decisão como a aprendizagem de árvore de decisão a partir de classes rotuladas nas tuplas de treinamento. A estrutura da árvore de decisão é semelhante a um fluxograma, onde cada nó interno (não-folha) indica um teste de atributo, cada ramo representa o resultado de um teste e cada nó da folha possui um rótulo de classe. O nó de nível mais superior é chamado de nó-raiz. A seguir,

[5] Chaffey 2006, pag. 378

a árvore de decisão que explicita se um cliente, de acodo com sua idade, irá efeturar ou não a compra de um computador:

Figura 2.9: Árvore de decisão



Para Ian e Frank (25), as árvores de decisão podem ser representadas por uma abordagem "dividir para conquistar" para resolução de problemas de aprendizagem, a partir de um conjunto de instâncias independentes. Os nós em uma árvore de decisão "testam" um atributo específico, comparando seu valor com uma constante. No entanto, algumas árvores podem comparar dois atributos com outros ou utilizarem uma função para tal. As árvores de decisão podem ser classificadas em dois tipos: árvores de regressão (regression trees), que são utilizadas para estimar atributos numéricos, e árvores de classificação (classification trees), usadas para análise de variáveis categóricas.

O algorítimo *C4.5* é considerado um exemplo clássico de método de indução de árvores de decisão. O *C4.5* (26) foi inspirado no algoritmo *ID3* (27), que produz árvores de decisão a partir de uma abordagem recursiva de particionamento de um conjunto de dados, utilizando conceitos e medidas da Teoria da Informação (28).

As árvores de decisão têm uma característica peculiar, a saída do modelo de predição (o output), com regras se – então é claramente perceptivel por analistas humanos. Essa qualidade é utilizada para interpretar os resultados.

2.5 Regressão Logística

Os modelos de regressão (linear e logística) são técnicas para analisar o relacionamento entre variáveis. No entanto, a regressão linear é utilizada para problemas de natureza contínua, sendo que a regressão logística é semelhante, contudo, a variável dependente não é contínua, é discreta ou categórica (29).

A regressão logística está definida como o logarítmo a seguir:

$$log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{2.3}$$

onde $\pi(x)$ é definido como:

$$\pi(x) = \frac{1}{1 + e^{-\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p}} \tag{2.4}$$

e $x_1, x_2, ..., x_p$ são as variáveis a serem exploradas.

A aplicação da regressão logística foi utilizada, pela primeira vez, com sucesso na oferta de crédito nos anos seguintes ao fim da 2ª guerra mundial, para tomar decisão de oferecer crédito a terceiros (30). A regressão logística é comumente aplicada para problemas de classificação binária (ou booleano).

3
Metodologia

A metodologia utilizada é um modelo preditivo de Suporte a Decisão que forneça informação suficiente a um gestor para decidir quando e por onde enviar uma frota de caminhões por determinada rodovia que apresente retenções crescentes de logística de cargas. As soluções disponíveis que existem tais como; Google Maps, Waze e outros dessa natureza somente exibem informações momentâneas, produzidas e compartilhadas pelos utilizadores dos aplicativos ou por informações provindas de GPS, contudo não analisam dados históricos dessas rodovias nem fazem predições futuras sobre o comportamento delas.

3.1 Plano geral da metodologia

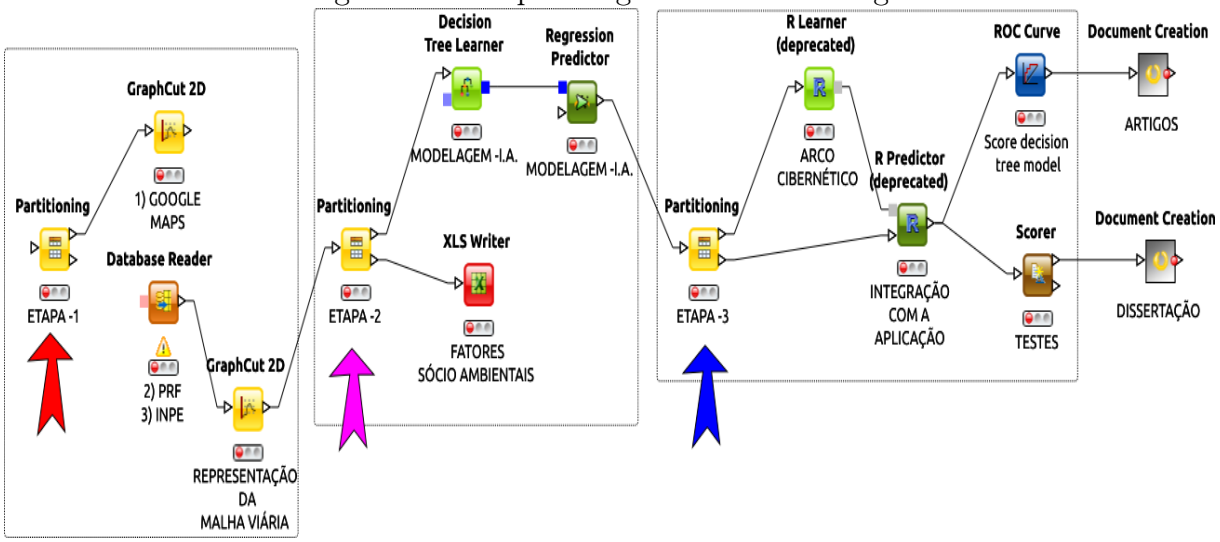
Propomos um plano que contemple 3 etapas, cada uma dividida em fases atinentes. A figura a seguir ilustra essa metodologia descrita graficamente, onde as três etapas são representadas por retângulos.

A etapa 1 contempla as fases da coleta das bases vetoriais, das bases de dados históricas e proposição de um modelo de representação único.

O retângulo central é a etapa 2, que consiste nas fases de Identificação dos fatores sócio ambientais na bases históricas, a modelagem do sistema preditivo e aplicação das técnicas de IA. Propomos inicialmente a Regressão Logística e Árvores de decisão, testes iniciais e depuração do modelo.

A última etapa da nossa proposta metodológica é a etapa 3, esta conterá um arco cibernético construído com os dados de redes sociais, como por exemplo a API do Twitter, este "per si" fará com que o modelo preditivo seja retro-alimentado, mantendo-o, ao longo do tempo atualizado na perspectiva do usuário gestor. Isso pois, modelos preditivos com o passar do tempo tendem a desfasar-se. Uma proposta algorítmica para substituir a API das redes sociais poderá ser desenvolvida e testada numa fase complementar, o algorítmo Ant-Miner poderá vir a ser um candidato de adaptação.

Figura 3.1: Etapas da gerais da metodologia



A representação da malha viária que acopla a estrutura dinâmica com a outra estrutura será o "front-end" da metodologia proposta estará na terceira etapa.

3.2 Proposição do modelo preditivo

O modelo preditivo foi construído utilizando bases de dados históricas da PRF (de acidentes e de paralisações ex: protestos) entre Janeiro de 2007 a Dezembro 2015. As bases de dados do Batalhão de Polícia de Rodoviária estadual – BPRv vieram entre Janeiro/2010 a Julho/2016, cortes em ambas as bases foram feitos para adequar as datas. Essas bases de dados são integradas gerando um único e complexo modelo preditivo que será acoplado a estrutura dinâmica.

3.3 Reflexão sobre as tecnologias utilizadas no modelo preditivo

Não existe uma técnica de mineração que generalize os mais diversos ambientes preditivos, mas sim um "pool" dessas técnicas onde uma complementa outra. As técnicas preditivas tradicionais que contemplam análise de grandes massas de dados como base heterogêneas são possíveis quando adaptadas para uma forma comparável à que foram inicialmente concebidas, por que as variáveis em uma base de dados a priori guardam pouca relação as variáveis de outra base de dados. neste caso essas variáveis ou são excluídas ou são transformadas a fim de "guardarem" um correlação com a outra base de dados. Quando há uma proximidade entre as bases de dados, na fase de transformação de dados, onde são criadas novas variáveis, a proximidade entre as bases se estreita. Nesta pesquisa, bases heterogêneas foram integralizadas num única grande base, onde as variáveis independentes foram em sua maioria preservadas e/ou construídas novas, nas bases onde não haviam correspondência, respeitando a lógica do negócio.
A tabela a seguir descreve as variáveis originais na base de dados de acidentes da PRF

3.3.1 Metodologia utilizada para coleta dos dados

As informações para suprir nosso modelo preditivo estão disponíveis na Internet, em sua maioria são Dados Governamentais Abertos, tais como os dados da PRF, INPE e IBGE. Isto são iniciativas governamentais para fomentar a participação popular, dentro outros motivos, essas informações são também conhecidas como *open data* (31), contudo os dados referentes à PRF e ao BPRv, para esta pesquisa, foram cedidos pelos respectivos órgãos governamentais (ver anexos) já em formato CSV para serem utilizados exclusivamente nesta pesquisa. Isso possibilitou ganho qualitativo nos dados evitando passar pelos transtornos como descreve Costa (2015) quando coletou os dados diretamente da Internet.(32) As bases de dados do INPE e do base de dados do IBGE apresentaram boa qualidade o que justificou serem serem coletados diretamente da Internet.

Tabela 3.1: Variáveis originais da base de acidentes

Ano	Ano da ocorrência do acidente
Mês	Mês de ocorrência do acidente
Num	Número do mês do acidente ex: 1 = Janeiro
KM	Numeração do quilômetro
BR	Numeração da Br
Latitude	Latitude da ocorrência
Longitude	Longitude da ocorrência
Condição Pista	Condição da pista: seca, molhado, ...
Restrição de Visibilidade	Restrição de visibilidade: inexistente, neblina, .., outros
Tipo Acidente	Tipo de Acidente: atropelamento, colisão lateral,..
Cauda Acidente	A possível causa do acidente: Falta de atenção, ...
Sentido Via	Sentido da via: crescente, decrescente
Traçado Via	Tipo de traçado da via: reta, curva, cruzamento, ...
Município	Localidade onde ocorreu
Tipo veículo	Tipo de veículo envolvido no acidente
Data Inversa	Data do acidente no formato dd/mm/aa
Horário	Hora que ocorreu o acidente no formato hh/mm/ss
Qtd Feridos Graves	Quantidade de feridos graves envolvidos
Qtd Feridos Leves	Quantidade de feridos leves envolvidos
Qtd Ilesos	Quantidade de ilesos envolvidos
Qtd Mortos	Quantidade de mortos envolvidos
Qtd Pessoas	Quantidade de pessoas envolvidos
Qtd Veículos	Quantidade de veículos envolvidos
Qtd Acidentes Graves	Quantidade de acidentes graves
Qtd Ocorrências	Quantidade de ocorrências

Na tabela seguinte; as variáveis originais da base de dados da PRF com interdições das vias (somente interdições que paralisaram as BRs, não contém acidentes, exemplo: passeatas, protestos)

Tabela 3.2: Variáveis originais da base de interdições

Comunicação	Código do agente que comunicou o incidente
Data Hora	Data hora no formato dd/mm/aa mm:ss
BR	Numeração da Br do incidente
KM	Numeração do quilômetro do incidente
Trecho	Local onde ocorreu o incidente

3.3.2 As variáveis do modelo preditivo

Algumas técnicas de IA são altamente sensíveis a dados ausentes os "missing data" à dados com pouca consistência e outros tipos de dados comuns em bases mantidas sem um bom critério de inserção dos dados. A variável dependente foi designada como **gargalo** e as variáveis independentes (ou explicativas) são:

Tabela 3.3: Variáveis do modelo preditivo

KM	Numeração do quilômetro
BR	Numeração da Br
condPista	Condição da pista: seca, molhado, ...
restVisibili	Restrição de visibilidade: inexistente, neblina, .., outros
tipoAcident	Tipo de Acidente: atropelamento, colisão, paralisação,...
tipoDano	Tipo de Dano: leve, médio, grave
Municipio	Localidade onde ocorreu
Ano	Ano que ocorreu o acidente
Mês	Mês que ocorreu o acidente
Dia	Dia que ocorreu o acidente
Hora	Hora que ocorreu o acidente

À base de dados da PRF, relativas a interdições das vias, por motivos diversos, não haviam variáveis tais como; visibilidade, condições da via, gravidade paralisação e outras. Foram incorporadas à essa base essas novas variáveis, para populá-las, adotou-se a lógica; presumivelmente protestos são realizados com boa visibilidade, em condições de via razoáveis e a gravidade da paralisação foi considerada leve.

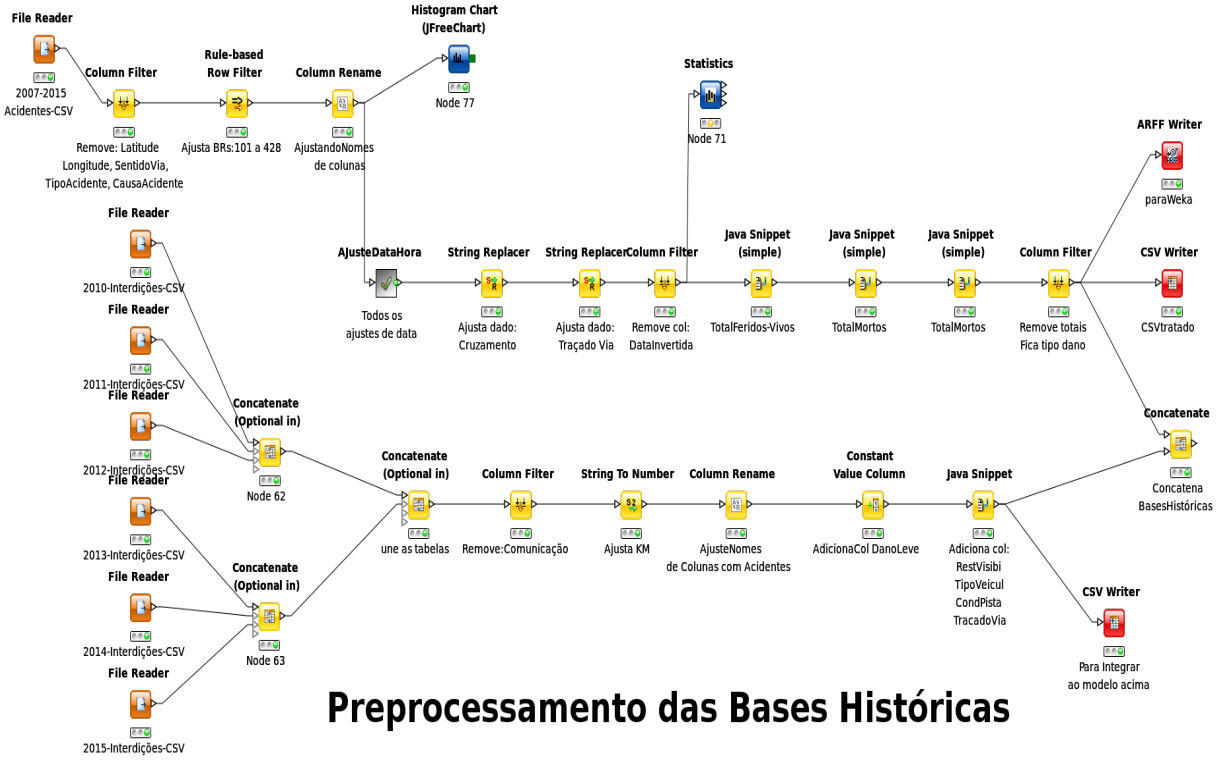
As técnicas como Redes Neurais Artificias (MLP) [CITAR], Árvores de decisão (CART) [CITAR], Regressão logística (MLR) [CITAR] fornecem visão generalizada dos fatores preponderantes, levantando padrões ocultos nos dados. Esta fase é conhecida como Aprendizagem de Máquina (acrônimo de Machine Learning)

a Redes Neurais Artificias do tipo *Multi Layer Perceptron* – (MLP) têm capacidade de receber várias entradas ao mesmo tempo e distribuí-las de maneira organizada, além são simples de implementar e trazem resultados satisfatórios em grandes bases de dados.

b Árvores de decisão para classificar acidentes do tipo *Classification and Regression Tree* – (CART) foi empregue por Pakgohar et al no artigo *The role of human factor in incident and severity of road crashes based on the CART and LR regression a data mining approach* com nível de acurácia próximo aos 80%

c Regressão logística tipo *Multinomial Logistic Regression* – (MLR) fornece a possibilidade de aprofundamento em vários níveis de busca sendo a mais apropriada, já que Regressão logística tradicional não permite aprofundamento desse tipo no espaço de busca.

Figura 3.2: Etapas 1 – Coleta e união das bases históricas de acidentes e interdições



3.4 Extração do conhecimento

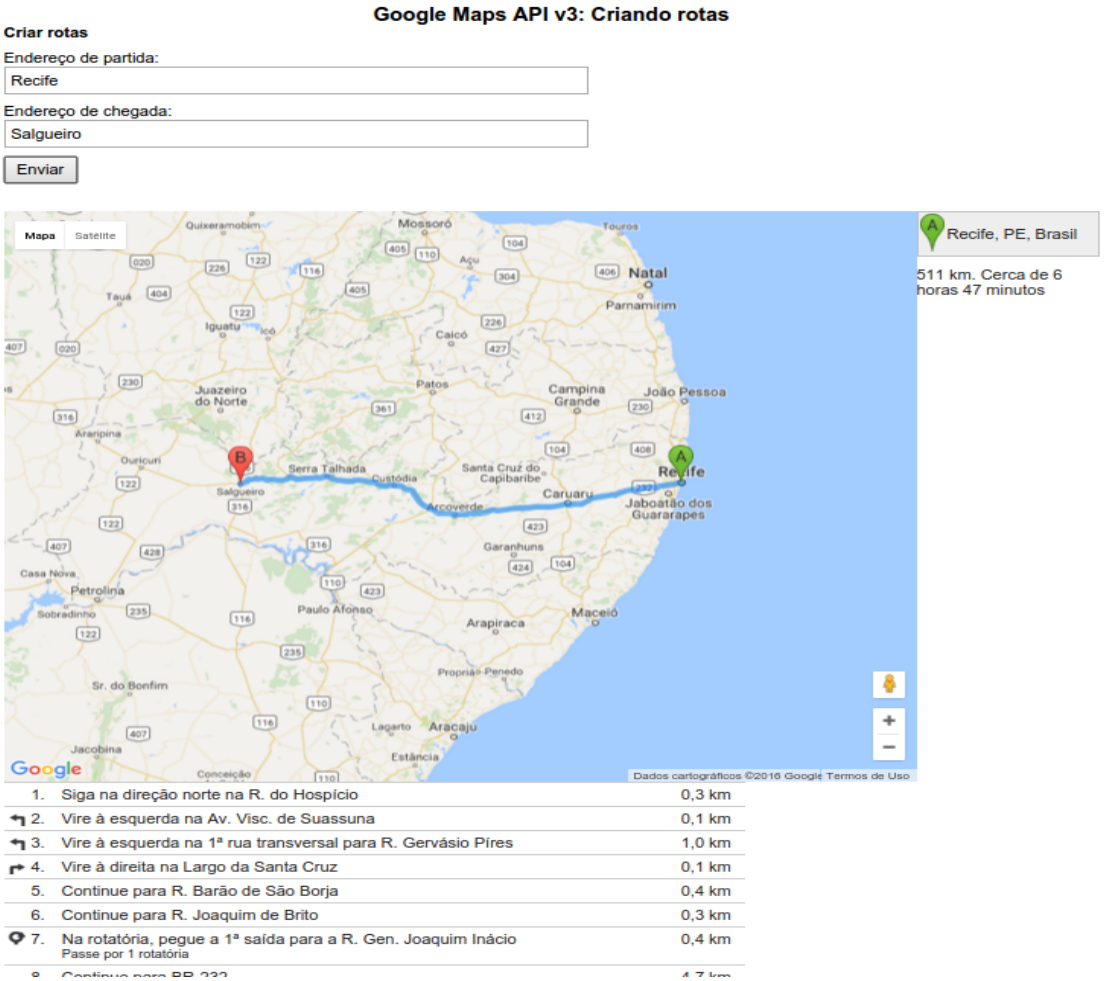
3.5 Acoplamento com a estrutura dinâmica

A estrutura dinâmica é composta por duas API's, uma disponibilizada pela Google, através do Google Maps que está atualmente na versão V3 e outra uma API do Twitter. A API do Google Maps proporciona uma "leitura" atualizada em forma de mapa no momento em que a estrutura dinâmica "roda".

A API do Twitter também tem a possibilidade de atualizar o modelo preditivo, contudo o objetivo desta é faz um Arco cibernético, retroalimentando todo o sistema com novas informações, pensamos que isso permite que o primeiro módulo (preditivo) seja atualizado de tempos em tempos, quando for
.

A API Google Maps portanto é o "front-end" do Sistema e uma futura aplicação que poderá ser desenvolvida para ser executada em um aparelho celular, "Smartphones", com capacidade para executar aplicativos mais complexos.

Figura 3.3: Etapas da metodologia



Concluída as três etapas e com as informações geradas pelo modelo serão escritos artigos científicos pertinentes à pesquisa em lide e a escrita da dissertação.

3.6 Cronograma

Tabela 3.4: Cronograma – 12 meses

Etapas/2016	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**	**Jan/17**
Rev. da literatura.	x	x	x	x	x	x	x	—	—	—	—	—
Etapa – 1	x	x	x	—	—	—	—	—	—	—	—	—
Etapa – 2	—	—	—	x	x	x	—	—	—	—	—	—
Etapa – 3	—	—	—	—	—	—	x	x	x	—	—	—
Escrita de artigos	—	—	—	—	—	—	—	—	x	x	x	—
Escrita da dissertação	—	—	—	—	—	—	—	—	x	x	x	—
Defesa	—	—	—	—	—	—	—	—	—	—	—	x

4
Resumo da Proposta, Discussão

4.1 Resumo da Proposta

As rodovias federais brasileiras que cruzam as regiões metropolitanas das grandes cidades se apresentarão sempre como um gargalo no fluxo do trasporte de cargas devido ao crescimento do número de veículos que nelas trafegam.

A solução proposta visa reduzir e dirimir os atuais gargalos burocráticos e tecnológicos para obtenção das bases de dados históricas de entidades públicas, os direitos autorais para utilização de APIs e de tecnologias específicas necessárias para apropriação via Internet. Entendemos ser normal esse cuidado por se tratar de informações de órgão que devem primar pelas informações de seus usuários. Mas tentaremos suprir demandas reais e importantes sem que isso represente algum risco à privacidade dos geradores de dados. Em suma, nossa proposta pretende mitigar o gargalo da logística de transporte de cargas, ofe- recendo uma solução possível à gestão de frotas de veículos que trafegam em rodovias, notadamente no caso do entrono metropolitano do Recife.

4.2 Discussão

Algumas propostas vêm sendo amplamente difundidas pelas mídias, tais como o arco metropolitano. Arcos metropolitanos, para além dos transtornos de se contruir um, são muito caros, requerem constantes manutenções e com o passar dos anos, com o crescimento populacional no seu entorno, tornam-se

novamente um novo gargalo para o transporte de cargas.

Gerir como as rodovias são utilizadas é a maneira mais racional, elas estão ai para auxiliar no trasporte de pessoas, mercadorias e para serviços, portanto é de todos e todos têm o dever de contribuir preservando-as e respeitando o direito dos outros.

Referências Bibliográficas

1 QUADROS, C. I. D. Dez Anos Depois do. *Intercom*, v. 1995, p. 65–69, 2005.

2 DILSIZIAN, L. E. S. E. Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized mecial diagnosis and treatment. p. 1.

3 BNDES. Perspectivas do investimento, n. 2, out. 2013. *Perpectivas do Investimento 2014-2017*, p. 2, 2013.

4 BITOUN, J. et al. Região Metropolitana do Recife no Contexto de Pernambuco no Censo 2010. p. 25, 2012. Disponível em: <http://www.observatoriodasmetropoles.net/download/Texto_BOLETIM_RECIFE_FINAL.pdf>.

5 IBGE, I. B. de Geografia e E. Região metropolitana do recife no contexto de pernambuco no censo 2010. 2014. Disponível em: <http://www.cidades.ibge.gov.br/painel/frota.php?codmun=261160&search=pernambuco/recife/infograficos:frota-municipal-de-veiculos/&lang=_ES>.

6 JAGADISH J. GEHRKE, A. L. Y. P. J. M. P. R. R. H. V.; SHAHABI, C. Exploring the inherent technical challenges in realizing the potential of big data. *Comunication of the ACM*, v. 57, n. 7, p. 86–96, July 2014.

7 DORSEY E. WILLIAMS, B. S. J.; GLASS, N. Twitter. July 2015. Disponível em: <https://pt.wikipedia.org/wiki/Twitter>.

8 FILHO, J. H. M. [S.l.: s.n.], 2012. 153–162 p. ISBN 978-85-7522-278-2.

9 DEAN, J.; GHEMAWAT, S. MapReduce : Simplified Data Processing on Large Clusters. *Communications of the ACM*, ACM, v. 51, n. 1, p. 1–13, 2008. ISSN 00010782. Disponível em: <http://portal.acm.org/citation.cfm?id=1327492>.

10 ARANHA, C.; PASSOS, E. *A Tecnologia de Mineração de Textos*. 2006. 1–8 p.

11 SRIVASTAVA, V. K. A.; SINGH, N. Review of decision tree algorithm: Big data analytics. *International Journal of Informative & Futuristic Research*, v. 2, n. 10, p. 3644–3654, 2015.

12 AMIN, A. et al. A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction. v. 353, p. 215–225, 2015. Disponível em: <http://link.springer.com/10.1007/978-3-319-16486-1>.

13 WIRTH, R. Crisp-dm 1.0 – step-by-step data mining guide. 2000.

14 EGAN, J. P. Signal detection theory and roc analysis. New York, USA: Academic Press, 1975.

15 ANAESTHETIST, T.

16 SOUZA, C. R.

17 AHMED, H.; GLASGOW, J. Swarm intelligence: concepts, models and applications. *School of Computing, Queen's University*, Citeseer, 2012.

18 CHAMBERS, W. D. Computer simulation of dental professionals as a moral community. *Medicine, Health Care and Philosophy*, Springer, v. 17, n. 3, p. 467–476, 2014.

19 BLUM, C. Ant colony optimization: Introduction and recent trends. v. 2, n. 4, p. 353–373, 2005. ISSN 15710645.

20 DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. *Theoretical Computer Science*, v. 344, n. 2-3, p. 243–278, 2005. ISSN 03043975.

21 BAIG, A. R.; SHAHZAD, W. A correlation-based ant miner for classification rule discovery. *Neural Computing and Applications*, v. 21, n. 2, p. 219–235, 2012. ISSN 09410643.

22 BARFORD, P.; YEGNESWARAN, V. An inside look at Botnets. *Malware Detection*, v. 27, p. 171–191, 2007. ISSN 03601315. Disponível em: <http://www.springerlink.com/index/10.1007/978-0-387-44599-1>.

23 GOOGLE, I.

24 HAN, J.; KAMBER, M. Data mining: Concepts and techniques. Elsevier, San Francisco, v. 2 edition, 2006.

25 WITTEN, I. H.; FRANK, E. Data mining: Practical machine learning tools and techniques. *Elsevier, San Francisco*, v. 2 edition, 2005.

26 QUINLAN, J. R.

27 QUINLAN, J. R. Discovering rules by induction from large collections of examples. *Expert systems in the micro electronic age. Edinburgh University Press*, In D. Michie, 1979.

28 HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.: s.n.], 2006.

29 WEST, D. Neural network credit scoring models – computers and operations research. p. 1131 – 1152, 2000.

30 M, H. J. Logistic regression models. Chapman and Hall – CRC Press, 2009.

31 2016. Disponível em: <http://dados.gov.br/dados-abertos/>.

32 COSTA, J. D. J.; BERNARDINI, F. C.; FILHO, J. V. A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. *AtoZ: novas práticas em informação e conhecimento*, v. 2, n. 2014, p. 1–26, 2015. Disponível em: <http://ojs.c3sl.ufpr.br/ojs/index.php/atoz/rt/printerFriendly/41346/25356>.