Minerando dados em grandes quantidades - *Big Data*: SURVEY

Othon L. T. Oliveira
Mestrando em Engenharia de Sistemas
Universidade de Pernambuco
Email: olto@ecomp.poli.br

Fernando B. L. Neto
Universidade de Pernambuco
PhD - UK
Email: fbln@ecomp.poli.br

***Resumo*—Este artigo pretende fazer uma explanação de um novo paradigma, o da informação ilimitada, conhecido como "Big Data" e propor uma solução para recuperação de informação para grandes espaços de buscas. Atrelado a este paradigma e reforçando o conceito de que uma nova onda está à frente da Internet, onde coisas se conectam entre si, conhecida como Internet das Coisas ou *Internet of Thigs* (IoT). A Internet das coisas é também conhecido como a "terceira onda da Internet"; onde computadores conectados a eletrodomésticos; telefones celulares ligados à semáforos, recebem informações sobre o tempo, acidentes e congestionamento e passam essas informação ao carro, que avisa o condutor o que está à frente. Conectar esses novos "atores" ou coisas que interagem com os atores humanos é necessário minerar as informações da Internet das coisas, transformando-os em dados para extrair-lhes informação relevante a fim de se aplicar na solução dos problemas do dia a dia.**

Palavras-chave: Big Data, Mapear e Reduzir, Hadoop, Enxame de robôs, Inteligência em enxames, Aprendizado em profundidade, Aprendizado de máquina, Otimização por enxame de partículas, Otimização por enxame de formigas, Pesquisa por cardume de peixes, Logística, Roteamento .

Abstract – This paper intends to make an explanation of a new paradigm, the paradigm of unlimited information known as "Big Data"and to propose a solution for information retrieval for large spaces searches. Coupled to this paradigm and, reinforcing the concept that a new wave lies ahead for the Internet, where things connect with things, we are not as well known Internet of Things or simply IoT. The IoT is also known as the "third wave of the Internet,"the acronym is confirmed because something like computers connected to home appliances, mobile phones connected to the traffic lights, getting weather information, accidents and congestion and to pass this information to the car, which warns the driver what's ahead. This is the reality that approaching and we intend to analyze in this article how the new players, things, interact with traditional actors humans to solve problems of day by day.

Keywords: Big Data, Map Reduce, Hadoop, Swarm Robotics, Swarm Intelligence, Deep Learning, Machine Learning, Particle Swarm Optimization, Ant Colony Optimization, Fish School Search, Logistic, Routing .

1. Introdução

Inferir sobre algum assunto agora poderá ser coisa do passado. Astrônomos atualizam suas descobertas numa base de dados disponíveis para outros utilizarem, as ciências biológicas agora tem tradição em depositar seus avanços científicos em repositórios públicos, redes sociais estão focadas na Web; Facebook, LinkedIn, Microsoft, Tweeter e Yahoo sobrevivem coletando informações e repassando-as as empresas de telemarketing, empresas de comércio eletrônico como Amazon, Submarino, Americanas.com, Magazine Luíza, utilizam essas informações para vender mais e melhor, artigos científicos dos mais variados assuntos, das mais variadas áreas alimentam, todos os dias, com milhões de informações os *Data Centers*, isso é o *Big Data*.

Para fazer frente a esse novo paradigma e extrair informações com eficiência novas abordagem algorítmica se faz necessário. As técnicas tradicionais de busca não são eficientes para resolver muitos problemas com grande complexidade, i.e. ordens de grandeza dantes inimagináveis, especialmente as que possuem complexidade exponenciais, geradas por explosões combinatórias, felizmente, muitos desses problemas não-triviais são eficientemente resolvidos por soluções naturais (2).

Para resolver a problemática da explosão combinatória das informações são desenvolvidas técnicas para reduzir o tempo de busca, tornando os custos computacionais mais aceitáveis, na medida que não exploram todas as possibilidades no caminho até a solução procurada, somente o caminho curto onde é mais provável de se encontrar a solução, a essa técnica dá-se o nome de heurística.

2. Objetivos

Este artigo "Survey" tem como objetivo a priori mapear as mais diversas tecnologias empregadas para analisar à problemática do "Big Data" e seus paradigmas apresentados atualmente, extrair informações e conhecimentos, analisando as técnicas de Inteligência Artificial mais empregadas nesse campo do conhecimento científico. Como objetivo a posteriori propor uma solução à logística de cargas aplicada à realidade brasileira, objetivando assim uma alternativa ao traçado de rotas determinísticas, inspirado em buscas no "Big Data".

Para executarmos este *Survey*, foram feitas coletas de dados entre os meses de

3. Plano de execução do Survey

Primeiramente definidas as seguintes etapas:

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

3.1. Detalhamento das etapas

A. Coleta dos dados
Para coletar os dados foram escolhidas sete (7) palavras chaves com relevância ao tema dentro do meio ambiente da pesquisa, destacamos:

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

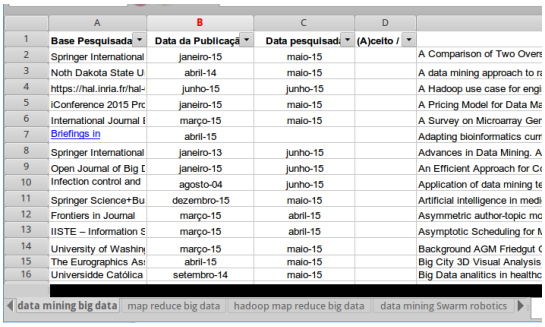
De posse das palavras-chave foram criadas planilhas, onde cada folha dessa planilha é representada por uma palavra chave. Incluiu-se no mínimo 30 artigos para cada planilha. Pretendeu-se, com essa arquitetura, construir rapidamente gráficos das mais diferentes matizes, tais como:

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

Uma das ferramentas utilizadas para extrair os dados referentes às palavras-chave foi o programa Mendeley, especializado em extrair dados de arquivos pdfs como os artigos, dissertações e teses. O Mendeley oferece uma opção para gerar dados extraídos em formato estruturado do tipo XML. A priori foi utilizado para gerar as planilhas, contudo, foi excluída essa opção por não trazer grandes ganhos, talvez se fossem criadas macros para tratar esses arquivos XML ficassem mais "limpos" pois o que mais foi relevante para esta fase do Survey formaram as colunas de cada planilha: A figura 1 mostra como está a planilha-Survey até este momento;

B. Seleção dos artigos
A seleção dos artigos foi primeiramente escolhida de forma qualitativa, pelas mais recentes publicações, dos últimos 5 anos, para ser mais exato entre 2010 e 2015. O segundo critério de seleção foi pelo órgão publicizador, haja vista os mais conhecidos tais como IEEE, Elsevier, Springer, e deles todos os jornais pertinentes, contudo outras fontes foram consideradas como universidades, seminários mais conhecidos na área.
Outro critério de seleção foi o quantitativo. Para isso foi estabelecido o 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 software para construtores de gráficos.

Figura 1: Planilha



C. Critérios de Inclusão/Exclusão
O critério de inclusão e exclusão foi baseado na leitura dos "abstracts" dos artigos. Devido a algumas palavras estarem muito na "moda" são citados em muitos artigos, apesar de conterem as palavras chaves, mesmo assim não traziam qualquer relevância para a pesquisa. Dessa forma foi criada uma pasta chamada "Rejeitados" para onde foram movidos esses artigos.
Outro critério de exclusão foi a data mais antiga, anteriormente à 2010, com alguma exceção dos artigos clássicos da área.

D. Leitura dos artigos
A leitura dos artigo iniciou-se tão logo terminaram os critérios de de Inclusão/Exclusão. Foram selecionados cerca de 100 artigos para serem lidos, não excetuando-se excluir mais algum que por ventura não tenham ficados "presos" na etapa anterior (C), dos filtros. Segue uma tabela com as datas do plano de execução

Tabela 1: Datas do Survey

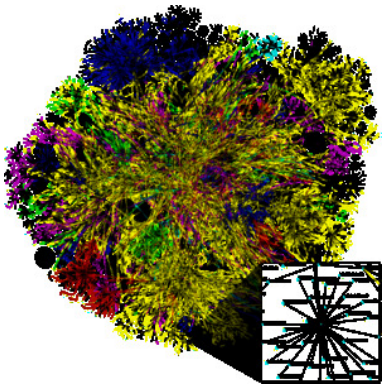
Data	A-Coleta	B-Seleção	C-Filtros
Abril a Junho	X	–	–
Junho e Julho	–	X	X
Agosto	–	–	X

4. A Internet

Uma *Internetwork* ou simplesmente internet é a conexão entre mais de uma rede e hoje em dia a maioria das redes se encontram conectadas. A Internet é um sistema de *internetwork* organizado e estruturado, a mais notável das internets, uma colaboração de mais de centenas de milhares de redes. Já "World Wide Web" (WWW) é apenas um dos muitos serviços que funcionam dentro da Internet.

O acesso à Internet está condicionado a Provedores de Acesso (ISP), que se classificam em internacionais, que conectam países; nacionais (*backbones* criadas e mantidas por empresas especializados) ; regionais ligadas a outros ISPs, normalmente têm taxas de transmissão menores; e locais que se conectam a ISP regionais ou nacionais e oferecem serviços de conexão a usuários finais.

Figura 2: Internet



4.1. Redes de dispositivos conectados

Uma rede é um conjunto de dispositivos (nós) conectados por links (caminho de transferência) de comunicação. Um nó pode ser um computador, uma impressora ou outro dispositivo de envio e/ou recepção de dados, que estejam conectados a outro nó da rede (1).

Por mais simples e menor que seja, três critérios são considerados os mais importantes que uma rede deve atender: Desempenho, que envolve a capacidade de vazão (throughput) e o atraso (delay); Confiabilidade, que envolve o tempo de recuperação quando falha e sua robustez caso haja alguma catástrofe; e Segurança, que envolve proteção ao acesso de dados e proteção contra danos e perdas (1) (3).

5. Big data

"Em 2010 empresas e usuários armazenaram mais de 13 exabytes de novos dados" (4).

O paradigma dos 5 V's exemplifica o Big data, este pode ser descrito como:

1 – Volume de dados;
2 – Velocidade para acessar esses dados;
3 – Variedade de informações;
4 – Veracidade nos dados encontrados;
5 – Valor atribuído aos dados.

Dentre as mais diversas ferramentas encontradas destacamos Map Reduce, da "framework" Hadoop, por ser de livre uso "freeware" e ter grande tolerância 'a falhas e fácil implementação, talvez explique a opção de baixo custo para a maioria das empresas e pesquisadores. A tecnologia congênere, desenvolvida pelo Google "Google File System" também tem seu destaque. Essas são as mais conhecidas dos "aventureiros" que estão desbravando o Big Data. Não será tarefa trivial inferir e desenvolver novas ferramentas para o Big Data pois este tem assumido, nos mais recentes anos proporções gigantescas como descrito na tabela 1

Tabela 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}

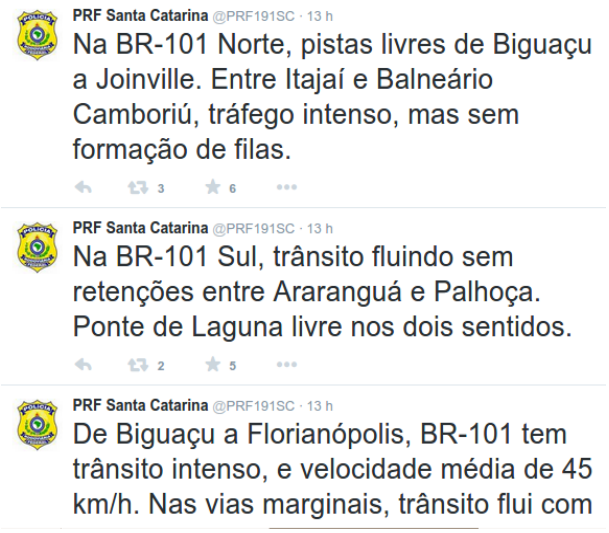
Atrelada a essa produção massiva de dados, uma nova onda está sendo vislumbrada, é chamada de "A terceira onda da Internet" , onde coisas se conectam com coisas; produtos nas gôndolas do supermercado com um novo tipo de etiqueta se conectam a uma leitora de radio frequência à alguns metros de distância, 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 expira a data de validade de determinados produtos, sem precisar abri-la, e quando acabarem esses produtos, a própria geladeira informará ao supermercado a falta deles, reservando o próximo rol de compras. Assim será essa onda de coisas conectadas, chamada de Internet das Coisas ou *Internet of Things* (IoT), que fará com que os dados no *Big Data* sofram explosão combinatória de informações multiplicando exponencialmente as dimensões deste.

As redes sociais são um arcabouço de informações sobre todo tipo de assunto vivenciado pelas 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. O Twitter é uma rede social, onde os usuários escrevem num pequeno espaço com cerca de 140 caracteres, os mais diversos assuntos. A ideia inicial do Twitter era que 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:

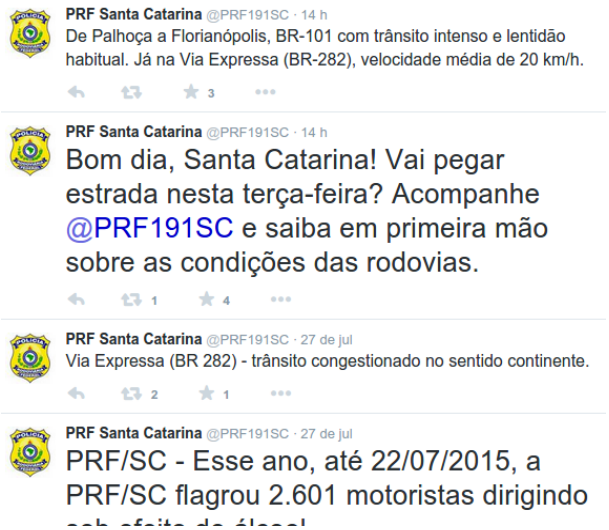
A Polícia Rodoviária Federal de Santa Catarina, disponibilizou às 13hs através do canal @PRF191SC, informações relevantes sobre o trânsito naquela localidade, num universo temporal variado, por exemplo: entre Itajaí e Balneário Camboriú o transito está intenso, isso sugere que a frota de

Figura 3: Twitter

(a) Twitte - 1



(b) Twitte - 2



caminhões que acompanhamos até este local deva ter uma rota alternativa, caso persista esta situação por muito tempo. No primeiro twitte da segunda coluna em Via Expressa (BR 282) trânsito lento com velocidade de 20km/h praticamente congestionado, novamente sugere que devamos "pensar" numa rota alternativa, caso esse 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, os wazers compartilham informações sobre o trânsito, em tempo real. Contudo as informações somente estão disponíveis no momento em que são postadas pelos utilizadores por um período de tempo pequeno, caso não hajam utilizadores trafegando pelas vias ou esses utilizadores não tenham disponibilidade em postar informações, não há o que se compartilhadar. Outra 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, o trânsito das grandes cidades têm disponíveis câmeras de monitoramento do trânsito nos semáforos, alguns com cobertura por canais de televisão, câmaras de segurança próximos às rodovias também coletam informações, tudo em tempo real. Os dados desses dispositívos geralente são gravados sendo conhecidos como *stream* de dados. Esses *streams* podem estar disponibilizados na Internet em sítios eletrônicos especialmente construídos para isso, como o "vejaaovivo" [1] e outros.

Os dados disponibilizados pelos diversos meios de comunicação não estão em formato que possam ser utilizados imediatamente, precisam antes serem processados. Esses dados não processados são conhecidos como "dados frios". O processo de tratar as informações, retirando-lhes o "lixo"; 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 foi criado um tipo arquitetura para computadores trabalharem em conjunto formando um *cluster* essa arquitetura é conhecida como *filesystem*[2].

O goolgle é o maior ator do *Big Data*, ele desenvolveu um modelo computacional para pesquisas na Web, e tem apresentado o uso eficiente da técnica MapReduce com modelos de programação combinados com tabelas conhecidas como BigTable. Introduziu o Google File System (8).

Outros grupos de pesquisadores desenvolveram Hadoop Distributed File System (HDFS) que é hoje o sistema de arquivos para Big Data mais utilizado.(10)

6. Hadoop – MapReduce

Hadoop MapReduce é uma técnica recente, especialmente, projetado para o processamento de grandes conjuntos de dados distribuídos. Hadoop nasceu do Apache. O Apache é um servidor Web, o tal como o Hadoop. Servidores Web são computadores especialmente dedicados a traduzir programas feitos para Internet em página da Internet as quais pessoas possam ler, quando conectam-se 'a Internet. MapReduce é um modelo de programação para expressar cálculo distribuído em quantidade maciça de dados e uma estrutura de execução para dados em larga escala e processamento em clusters de servidores. Foi originalmente desenvolvido pela Google e construído sobre o bem-conhecido princípios em paralelo e processamento distribuído. O Hadoop é a implementação de código aberto do MapReduce escrito em java que fornece, tolerância a falhas, escalável e confiável

1. *http://vejoaovivo.com.br* sítio eletrônico onde encontram-se imagens de câmeras de trânsito em tempo real

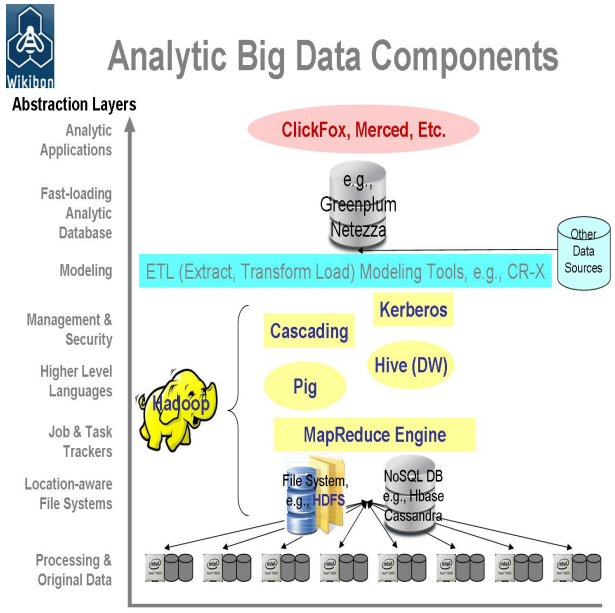
2. *Filesystem*. ou sistema de arquivos, referem-se à forma como os dados são armazenados, organizados e acessados, pelo sistema operacional, em cada partição no disco (ou no disco inteiro)

técnica de computação distribuída. O configurar o ambiente Hadoop envolve um grande número de parâmetros que são essenciais para alcançar um excelente desempenho. Ele permite que desenvolveres de aplicações distribuídas sem qualquer conhecimento possam programar computadores. Um ”Valor” e uma ”chave” formam um par de dados, isso é a estrutura básica de dados do MapReduce. Chaves e valores podem ser da forma de dados primitivos; como inteiros, ponto flutuante, e bytes brutos ou podem ser estruturas arbitrárias e complexas (listas, tuplas matriz associativa, e outras)

6.1. Hadoop - Map Reduce - Big Data

O *Big Data*, devido à sua natureza, foi definido como o paradigma dos 5 V's entre as grandezas: **Volume**, **Velocidade**, **Variedade**, **Veracidade** e **Valor**. O roteiro seguido pela informação desde o local onde é produzida até onde possa ser utilizada é exemplificado na imagem a seguir:

Figura 4: Big Data e Arquitetura Haddop



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* é o roteiro seguido pela informação para ser trasformada em dados “quentes”. A tecnologia utilizada, para tratar os dados poderá ser o Apache Hadoop ou similar, devido ao baixo custo de implementação e por ser uma tecnologia aberta, de livre utilização, conhecida como *Open Source*. O Haddop é uma arquitetura de milhares de computadores interligados e espalhados e pela Internet. Esses computadores são especializados em extrair dados do *Big Data* e transformá-los em dados relevantes (quentes). (9)

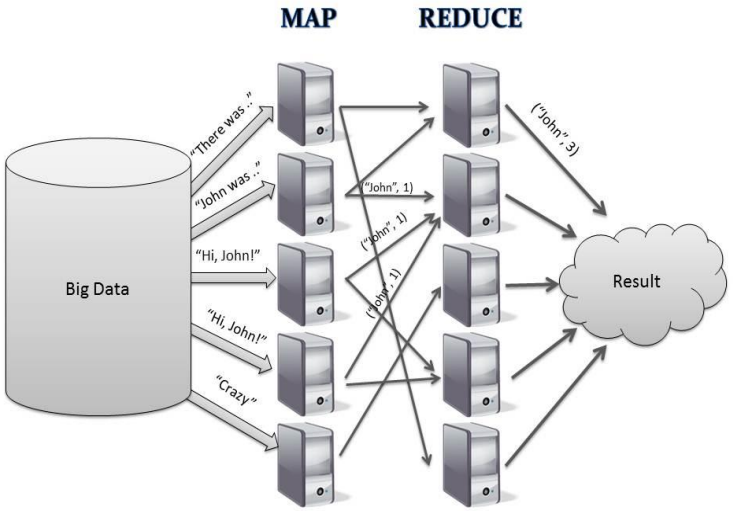
O Haddop funciona com um agrupamento em paralelo desses computadores, conhecido como *cluster*. Esse *cluster* tem a característica de grande escalabilidade; em torno de 3 000 computadores na operação de Map, dependendo da construção, e tolerância a falhas; quando um computador do *cluster* fica inoperante “cai”, os dados são salvos em outro computador. A estrutura de diretórios do Hadoop é conhecida como *Hadoop File System* (HDFS). O HDFS foi especialmente construído para lidar com as características descritas anteriormente.

6.2. Map Reduce - Big Data

O Map-Reduce é uma técnica conhecida desde a linguagem Lisp, que permite Mapear (Map) um conjunto de informações como palavras, imagens, paginas na Internet através do agrupamentos de milhares de computadores conhecido como *cluster* e mesclar essas informação conhecida como reduzir (Reduce) conforme um par de informações { chave:valor } . O Hadoop implementa o paradigma MapReduce

A imagem a seguir exemplifica a técnica Map-Reduce:

Figura 5: Técnica Map-Reduce



Durante a operação de “Map” podem estar envolvidos *clusters* de até 3.000 computadores “espalhados” pela Internet, caso seja utilizada a infraestrutura Hadoop sem virtualização. Na etapa de “Recude” o *cluster* cai consideravelmente, contudo o consumo de energia nessas duas etapas consideravelmente alto. (9) Este modelo de Map-Reduce, é um modelo de computação paralelizada proposto pelo Google para ser utilizado na Internet (11).

Muitas técnicas tem sido utilizadas para executar a operação de MAP, até virtualização de máquinas, contudo o grande poder do MapReduce está no *cluster* que é implementado com arquitetura em nuvem *Cloud Computing* onde os computadores com *File System* HDFS se comunicam entre si para formar um agrupamento.

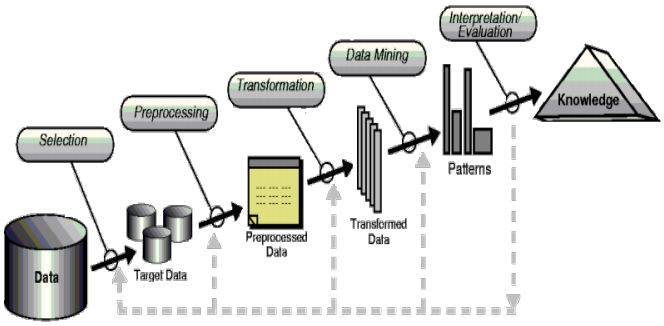
{aqui tirar do artigo: A Modified Key Partitioning for BigData Using MapReduce in Hadoop}

7. Data Mining

Técnicas de mineração de dados trabalham com dados estruturados, para poder extrair informações relevantes. Um dos maiores problemas na extração de informações são os "missing data" ou dados ausentes. Para contornar os dados ausentes exitem varias técnicas como preenchimento dos dados através de técnicas de inteligencia 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 6: Minerando dados no Big Data



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

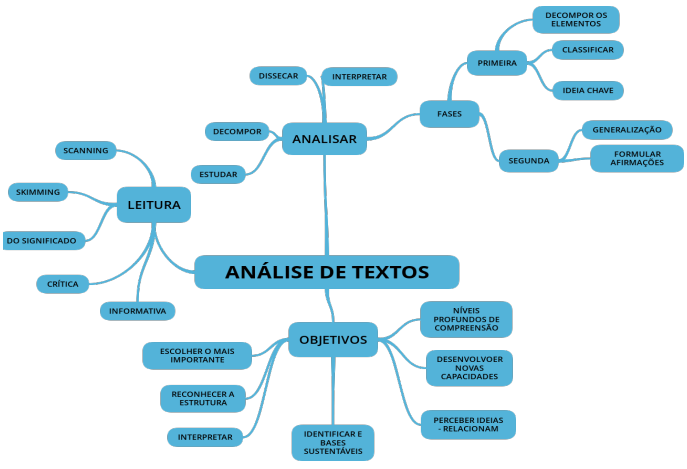
O Big data está representado, na imagem, onde se lê "Data", repleto em dados ausentes e/ou inconsistentes por isso conhecidos como dados não estruturados. Os balão onde se lê "Selection" representa a coleta das informações, para este artigo representa a seleção dos dados no Big Data vindos das mais diversas fontes, tais como, redes sociais, câmaras de trânsito, informações de satélites meteorológicos e muitas outras fontes. Armazenar qualquer quantidade de dados nessa etapa pode ser um grande problema, devido a quantidade, como mencionado anteriormente, porém os dados relevantes podem ser armazenados em "Target Data" com Hadoop e as técnicas de "Map" e "Reduce" poder-se-ia agrupar (clustering) informações ou ler os fluxos de dados (stream data) algumas técnicas de IA podem ser aplicadas nessa etapa, como "Data Mininng Swarm Robotics" através de Botnets e "Swarm Intelligence". No balão "Preprocessing" ainda são dados não-estruturados, para estruturá-los é preciso técnicas linguísticas pois existe lógica entre eles (12). Esses dados normalmente são coletados por técnicas de Mineração de Textos ou Mineração de Dados em Textos, mais uma vez técnicas de IA como "Machine Learning" como será descrito mais adiante. Uma vez transformados os dados em "Transformation" esses dados, já estruturados podem ser armazenados em Bancos de Dados conhecidos como Datawarehouse a exemplo do Hive, descrito na Figura 4. O processo de Mineração dos dados começa no balão "Data Mining". Nessa etapa são aplicados outros técnicas de IA como "Decion Tree", "Artificial Neural Network", "Logistic Regression", "Deep Learning" e muitas outras técnicas de classificação e extração de padrões. Todas essas etapas, na mineração de dados são recorrentes como indicam as setas pontilhadas. Algumas técnicas de mineração de dados são fortemente influenciadas pelas informações à entrada, como as Árvores de decisão ou "Decision Tree" (21) e Redes Neurais para minerar dados teriam milhares de neurônios na camada intermediária o que inviabilizaria a técnica aplicada a este contexto, portanto utilizar essas duas técnicas, para extrair informações do Big data, pode levar a inconsistências incontornáveis.

Utilizar técnicas de mineração de dados é para além de extrair dados, extrair conhecimento do negócio que se está analisando e com isso poder predizer os resultados futuros à saída do modelo, aquando determinados dados à entrada ocorrem. (13), essa técnica de extração de conhecimento chama-se *Knowledge Discovery Databases* (KDD).

7.1. Data Mining - Big Data

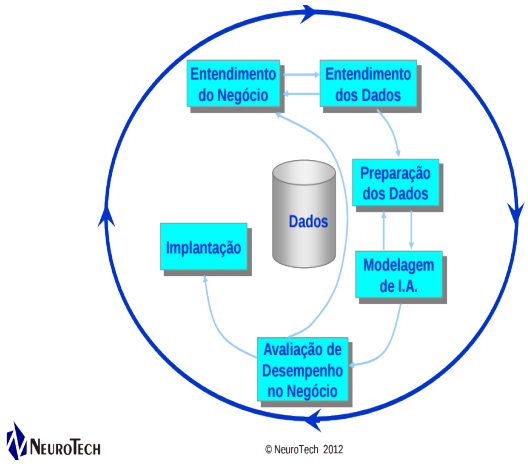
Minerar dados no Big data pode não ser uma tarefa atômica, devendo ser divida entre vários tarefas com vários processos em cada uma delas, como descrito anteriormente, pois os dados coletados até chegar a etapa de análise precisam ser tratados. Extrair conhecimento dos dados não processados não faz sentido, tratá-los só "per si" exige muito trabalho de IA como Mineração de dados em textos. Mineração em textos é inspirado em técnicas de "Machine Learning" (12). 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 7: Mapa mental de análise de textos



Existem diversos técnicas algorítmicas para isso. No entanto todas elas necessitam que os dados sejam checados e validados constantemente, essa técnica é conhecida como "CRoss Indrustry Standard Process for Data Mining" (CrIP-DM). A figura a seguir exemplifica.

Figura 8: Minerando dados no Big Data

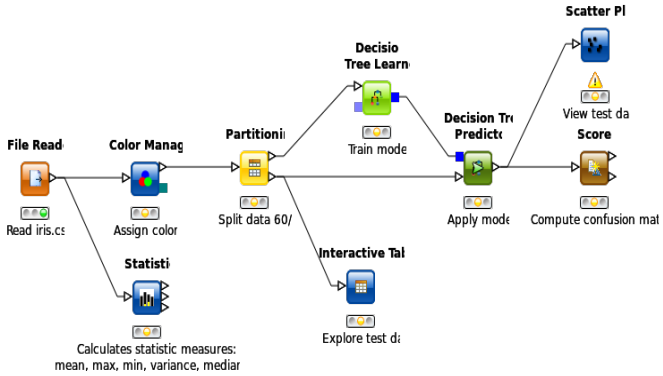


O **Entendimento do negócio** é uma fase crucial da mineração, onde um especialista (ou muitos) deve ser consultado, pois o analista de dados geralmente consegue fazer re-uso de conhecimento lendo periódicos e artigos, mas o especialista no negócio é um profissional essencial e deve estar com anos de experiência no negócio. Passada essa fase de Entendimento do Negócio o analista de dados poderá passar a fase do **Entendimento dos dados**, nesta fase o analista de dados "olha" para os dados com acurácia procurando identificar a qualidade dos dados. Dados ausentes são comuns em bancos de dados não estruturados, os"missing data" são sempre um problema a ser considerado, pode consumir muito tempo do analista de dados. Após superar a fase do entendimento do negócio o analista de dados passará a fase da **Preparação dos dados**, esta fase cobre a construção final do conjunto de dados, preparar os dados significa criar selecionar atributos criar tabelas 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 que exijam apoio a decisão uma tecnologia inadequada pode levar a decisões erradas, é comum retornar as fases anteriores para se adequar 'as técnicas aos dados, por exemplo um modelo de regressão logística para problemas binários andam juntos na maioria dos sistemas, redes neurais andam juntos com problemas de classificação e assim por diante, contudo sempre há espaço para novas tecnologias e sobre tudo hibridização nas tenologias. ver figura a seguir:

Um conjunto de treinamento de ser preparado na fase anterior, para se fazer testes aos algorítimos dos Modelos de IA, contudo isso somente serve para se medir o acerto e a acurácia do modelo utilizado.

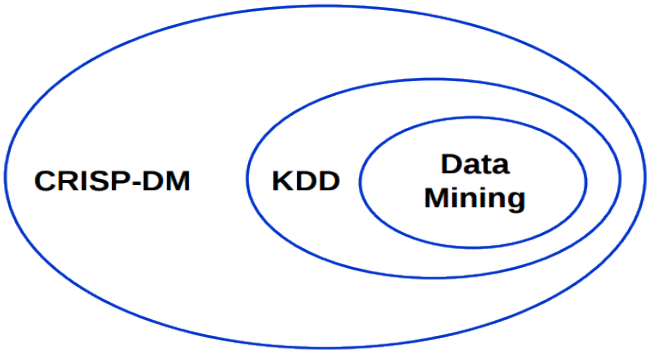
Métricas deverão ser propostas na fase de avaliação de desempenho pois minerar dados com poucos registros (¡

Figura 9: Modelo de IA baseado em "Decision Tree" para o Big Data



10.000) não costumam construir bons sistemas suporte a decisão baseados em de mineração de dados. **Avaliação de desempenho** em mineração de dados um ou muitos modelos devem ter sido construídos e testados e com alta qualidade da perspectiva da analise dos dados. Criar um modelo geralmente não é o fim do negócio, contudo é criar um modelo de entendimento do negócio. Até que se obtenha as respostas satisfatórias o modelo deverá ser refeito várias vezes até sua **Implantação**. A imagem a seguir descreve o domínio das técnicas aplicadas à mineração de dados:

Figura 10: Domínio das técnicas aplicadas a mineração de dados



8. Meta-heurísticas

As meta-heurísticas são soluções encontradas nas mais diversas espécies de seres vivos, por exemplo, as formigas quando vão em busca de alimentos, facilmente encontram um caminho mais curto entre o ninho e a fonte de alimentos, cardumes de peixes executam movimentos aparentemente aleatórios mas, quando em grupo, são precisos para fuga dos predadores, bandos de pássaros quando em busca de novos locais para ninhos ou de alimento, inspiram os mais diversos algoritmos inteligentes baseados em populações de animais sociáveis, demonstrando que há uma inteligência coletiva

nessas populações, desenvolvida ao longo do tempo e das interações entre essas espécies sociáveis e o meio ambiente. Essa classe de algoritmos, metaforiza o comportamento de tais populações, e promoveu o desenvolvimento de uma área que hoje é conhecida como computação bio inspirada ou computação natural. Essa área investiga a relação entre a computação e a biologia (e mesmo a sociologia), estudando soluções de buscas e otimização, modelando problemas mais eficientemente, baseado nas elegantes soluções encontradas pela natureza.

9. Enxame de partículas

Em 1989, G. Beni e J. Wang cunharam a expressão *Swarm Intelligence*, no seu trabalho em Robotic Swarm (5). O estudo do reino animal aprofundou-se no estudo comportamental e possibilitou o melhor entendimento de como cooperam indivíduos dentro de um grupo e quais os mecanismos usados para controlar o enxame e condicionar o indivíduo, tais como a estigmergia. Por enxame, pode entender-se 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, e.g., um enxame de pássaros. Os 5 princípios da inteligência de enxame segundo Chambers (22), são:

- Proximidade: os agentes têm que ser capaz 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 as variações ambientais

Swarm Optimization - (PSO)

Kennedy e Eberhart (1995), criaram a popular Otimização por Enxame de Partículas do inglês *Particle Swarm Optimization* - (PSO), que na verdade é uma técnica de otimização de funções não-lineares baseado em populações. Foi inspirado no comportamento social em bando de pássaros, essa técnica é uma das mais conhecidas e investigadas hoje (15). Esse algoritmo (e todos os outros) têm um fator de convergência, para fazer com que encontre mais rapidamente a resposta procurada. Para acontecer isso foi introduzido um "poleiro" virtual. Com essa simples abordagem foi inaugurado uma nova família de algorítimos baseados em enxames.

No PSO, a população é chamada enxame e os indivíduos, partículas. Cada partícula se move no espaço de busca, à procura de regiões promissoras; cada partícula dessas representa uma solução candidata a resolver nosso problema. A equação utilizada para encontrar uma partícula no espaço de busca foi emprestada da cinemática:

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{1}$$

Onde $x_1(t)$ é a posição da partícula num determinado momento "t". O $x_i(t+1)$ é a posição atual da partícula.

A velocidade da partícula é de acordo com a equação:

$$v_i(t+1) = v_i(t) + Nbest + Lbest \tag{2}$$

$Nbest = c_1 r_1 j(t)[Nbest - x_i(t)]$
$Lbest = c_2 r_2 j(t)[Lbest - x_i(t)]$

Onde $v_i(t)$ é a velocidade num momento "t" qualquer e $v_i(t+1)$ é a velocidade atual da partícula. Os coeficientes $c_1 r_1 j(t)$ e $c_2 r_2 j$ são números que variam entre 0 e 1 para mudar a posição da partícula. O coeficiente $Nbest - x_i(t)$ é a melhor posição da partícula na vizinhança e $Lbest - x_i(t)$ é a melhor posição numa região que engloba essa vizinhança.

Fish School Search - (FSS)

Para contornar o problema explosão combinatória Carmelo e Buarque propuseram a meta-heurísticas da busca por cardume de peixes ou o *Fish School Search* - (FSS) (18).

Na busca FSS, cada peixe representa uma possível solução do problema. Em busca por enxame de partículas há o problema da degradação do enxame, quando aparentemente as partículas encontram um mínimo local (poderia ser máximo - depende da natureza do problema) "pensando" terem encontrado o mínimo global. Para contornar esse problema da degradação do enxame introduz-se operadores que façam com que o exame saia desses "fossos" de busca. O FSS possui operadores para evitar o problema da perda de qualidade, dentre esses podemos citar o operador de volatilidade, que faz com que o enxame expanda quando o enxame se concentra por muito tempo.

A equação que faz isso é a seguinte:

$$Bari(t) = \frac{\sum_{i=1}^{N} x_i(t) W_1(t)}{\sum_{i=1}^{N} x_i(t)} \tag{3}$$

Ant Colony Optimization - (ACO)

A optimização por colônia de formigas ou *Ant Colony Optimization* - (ACO) é uma técnica de otimização que foi introduzida desde os anos 90's (16) baseado no comportamento forrageiro de colônia de formigas. O comportamento forrageiro em diversas espécies (17) 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 essa 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 chamado de feromônio. Esse feromônio evapora com o passar do tempo, sendo assim, a medida que mais formigas sigam um determinado caminho, mais intenso o feromônio se fará presente.

A equação da evaporação do feromônio no ambiente é segundo a fórmula:

$$p(i,j) = \frac{[\tau(i,j)]^{\alpha}.[\eta(i,j)]^{\beta}}{\sum[\tau(i,j)]^{\alpha}.[\eta(i,j)]^{\beta}} \quad (4)$$

9.1. Data Mining - Swarm Intelligence

Recentemente algorítimos de classificação baseados em "Ant Colony Optimization" (ACO) tem sido experimentado em mineração de dados(19) 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 termos 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 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 a qualidade da regra. Quando todas as formigas construírem as regras delas as melhores regras são selecionadas, colocadas numa lista de regras descobertas. A amostra de treinamento corretamente classificada pela regra são removidas do conjunto de treinamento. Esse processo é continuado até que o número de amostras não cobertas é pequeno e o usuário estabeleça um limiar. O produto final é uma lista de regras ordenadas que será usada para classificar um conjunto de testes. A seguir o algoritmo de classificação AntMiner:

Algoritmo 1: ANT-MINER

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

início
  enquanto $conjTreino > MaxNaoDescoberto$ **faça**
    $t = 1$ /*índice de formmigas*/
    $j = 1$ /*índice de convergência*/
    /*inicializa todas as trilhas com algum feromônio*/
    repita
      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; */
      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*/*
      então
        $j = j + 1$;
        senão
          $j = 1$;
        fim
      fim
      $t = t + 1$;
    até $t >= No_ants$ *ou* $j >= NoRegrasConverg$;
    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}};
  fim
fim

9.2. Data Mining - Swarm Robotics

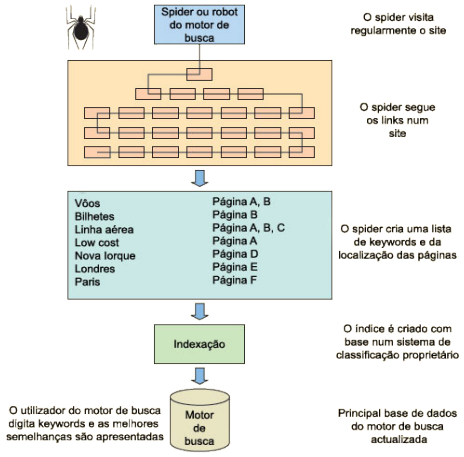
O rápido crescimento da Internet, tem trazido, a reboque, o contínuo crescimento da insegurança nos computadores e sistemas atuais (14). Ataques utilizando a computadores através de robôs conhecidos como Botnet é um problema contante para que lida com segurança da informação. No entanto os Botnet podem 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" página e indexando-as para que as buscas do motor de buscas da Google seja mais eficiente. A figura a seguir podemos ver um exemplo desses robôs: (20)

A utilização desses robôs é uma alternativa eficiente para coletar informações no Big data (ver "Selection" Figura 6) já que sua especialidade é essa. 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 6.1 (Hadoop – Map Reduce – Big Data).

9.3. Machine Learning

O desafio *Big Data* é resumido por alguns autoes a 3 V's (Velocidade, Variedade e Volume) dado que as informçãoes já tiveram um tratamento adequado (eliminando os últimos V's) contudo Velocidade, Variedade e Volume pode ser um fator determinístico na escolha da ferramenta mais adequada para se analisar dados do *Big Data* e extrair informação. As árvores de decisão são algorimos rápidos, contudo dados impuros podem comprometer o desepenho desse algoritmo. A fase de extração dos dados do *Big Data* são fortemente influenciáveis pelas variáveis escolhidas, (21) isso pode representar o desafio maior para implementar esta técnica, na figura a seguir exemplificamos essa tarefa:

Figura 11: Minerando textos no Big Data com robôs



10. Resultados esperados

Este Survey têm diversos propósitos preliminares e também conclusivos, para balizar pesquisas na área, apontando um caminho mais promissor relacionado ao assunto escolhido, o Big Dada. Um dos propósito é servir de referência científica a qualquer de busca por conhecimento científico, mostrando o trabalho que o pesquisador deve ter, o cuidado ao procurar esse conhecimento, sem contaminar-se na "floresta" de artigos científicos, dissertações e teses ofertados e, disponíveis hodiernamente na Internet.

Após a conclusão deste Survey, pretende-se incluir, os dados encontrados, no capítulo 2 da dissertação, servindo também de exemplo para todo mestrando que se aventurar na busca "indiscriminada" por artigos, deixando de lado uma parte importante do seu trabalho que são os dados estatísticos dessa fase e que pode ter assim uma abordagem científica já na fase inicial. A fase conclusiva, desse Survey, será a certeza de que o melhor caminho escolhido para o desenvolvimento das outras fases da dissertação foi baseado em métodos científicos, desde sua fase inicial, ficando disponíveis para outros pesquisadores o "como" deve ser todo trabalho de busca por artigos científicos.

11. Conclusão

Devido à natureza da problemática ser as dimensões, o Big Data tem tornado-se um desfio hérculeo para quem envereda-se a desvendá-lo. Inferir sobre o Big Data é o que muitos pesquisadores têm feito recentemente, utilizando-se para isso das mais diferentes tecnologias a disposição; Hadoop ou *Map Reduce* ou uma combinação dos dois, essas fronteiras vêm sendo desmistificadas aos poucos, mas ao mesmo tempo, novos desafios se fazem presentes, com a chegada da Internet das Coisas, onde tudo está conectado (23) desde eletrodomésticos (24) a carros, sinais de trânsito, fazendo com que os *smartphones* seja a ferramenta para se ter acesso a isso tudo. (6) Portanto estamos só no começo, no que está por vir, fazendo com que haja cada vez mais trabalhos nessa área e inspirações para novos algoritmos.

Acknowledgments

The authors would like to thank...

Referências

1 FOROUZAN, Behrouz A. Comunicação de Dados e Redes de Computadores. São Paulo: McGraw-Hill, 2008

2 Talles Henrique De Medeiros, Luís Fabrício Wanderley Góes, M. B.-B. C. E. I. M. (n.d.). *Computação Bioinspirada aplicada à Robótica*

3 E. SPECIALSKI (1999). *Gerência de redes de computadores e telecomunicações*

4 H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakontantinou, J. M. Patel, R. Ramakrishnan and C. Shahabi *Exploring the inherent technical challenges in realizing the potential of Big Data*, journal:Comunication of the ACM, volume=57, numbe=r7, pages=86–96, month=July, year=2014

5 H. Ahmed, and J. Glasgow, *Swarm intelligence: concepts, models and applications*, School of Computing, Queen's University, Citeseer, 2012.

6 SINGER, Talyta. TUDO CONECTADO: CONCEITOS E REPRESENTAÇÕES DA INTERNET DAS COISAS. 2012. Acessado em: 23 abril. 2015. Singer

7 Dorsey, J. Williams, B. Stone, E. and Glass, N. Acessado em Julho de 2015 Twitter

8 Filho, João Heriberto Mota, booktitle = Descobrindo o Linux: entenda o sistema operacional GNU/Linux, isbn = 978-85-7522-278-2, pages = 153–162, year = 2012

9 Conejero, Javier and Rana, Omer and Burnap, Peter and Morgan, Jeffrey and Caminero, Blanca and Carrión, Carmen, title = Analysing Hadoop power consumption and impact on application QoS, issn = 0167739X, journal = Future Generation Computer Systems, mendeley-groups = HadoopMapreduceinBigData, doi = 10.1016/j.future.2015.03.009, year = 2015 HadoopMapreduce

10 Lange, Benoit and Nguyen, Toan title = A Hadoop use case for engineering data, mendeley-groups = DataMiningBigData, year = 2015

11 Dean, Jeffrey and Ghemawat, Sanjay institution = Google, Inc., issn = 00010782, journal = Communications of the ACM, number = 1, pages = 1–13, pmid = 11687618, publisher = ACM, series = SIGMOD '07, title = MapReduce : Simplified Data Processing on Large Clusters, volume = 51, year = 2008 MapReduce

12 Aranha, Christian and Passos, Emmanuel, A Tecnologia de Mineração de Textos, booktitle = RESI-Revista Eletrônica de Sistemas de Informações, doi = 10.5329/171, issn = 1677-3071, keywords = Data minig,Intelligent information systems, mendeley-groups = Mineração Textos, number = 2, pages = 1–8, volume = 2, year = 2006

13 Amin, Adnan and Faisal, Rahim and Imtiaz, Ali and Changez, Khan and Anwar, Sajid, title = A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction, doi = 10.1007/978-3-319-16486-1, isbn = 978-3-319-16485-4, keywords = big data,stock prediction,text mining, mendeley-groups = DataMiningBigData, pages = 215–225, volume = 353, year = 2015, stockprediction

14 Barford, Paul and Yegneswaran, Vinod, title = An inside look at Botnets, doi = 10.1007/978-0-387-44599-1, isbn = 978-0-387-32720-4, issn = 03601315, journal = Malware Detection, pages = 171–191, volume = 27, year = 2007 Botnet

15 J. Kennedy and R. Eberhart (1995). *Particle swarm optimization. In Neural Networks, 1995. Proceedings.* IEEE International Conference on (Vol. 4, pp. 1942–1948 vol.4). http://doi.org/10.1109/ICNN.1995.488968

16 C. Blum *Ant colony optimization* booktitle: Physics of Life Reviews, title: Ant colony optimization: Introduction and recent trends, doi: 10.1016/j.plrev.2005.10.001, issn: 15710645, keywords: Ant colony optimization,Discrete optimization,Hybridization, number: 4, pages: 353–373, volume: 2, year 2005

17 M. Dorigo and C. Blum, *Ant colony optimization theory: A survey*, doi: 10.1016/j.tcs.2005.05.020, isbn: 0304-3975, issn: 03043975, journal: Theoretical Computer Science, keywords: Ant colony optimization, Approximate algorithms, Combinatorial optimization, Convergence, Metaheuristics, Model-based search, Stochastic gradient descent, number: 2-3, pages: 243–278, volume: 344, year: 2005

18 Filho, Carmelo J A Bastos and Neto, Fernando B De Lima and Lins, Anthony J C C and Nascimento, Antônio I S and Lima, Marília P. booktitle: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, *A novel search algorithm based on fish school behavior*, doi: 10.1109/ICSMC.2008.4811695, isbn: 978-1-4244-2383-5, issn: 1062922X, keywords: Fish school,Search algorithms,Social behaviour,Swarm intelligence, pages: 2646–2651, year: 2008

19 Baig, Abdul Rauf and Shahzad, Waseem, doi = 10.1007/s00521-010-0490-5, title = A correlation-based ant miner for classification rule discovery, issn = 09410643, journal = Neural Computing and Applications, keywords = Ant colony optimization (ACO),Classification rules,Data mining,Swarm intelligence, mendeley-groups = DataMiningSwarnIntelligence, number = 2, pages = 219–235, volume = 21, year = 2012

20 Fonte: Chaffey, page = 378, year=2006

21 A. Srivastava, V. Katiyar and N. Singh – Review of Decision Tree Algorithm: Big Data Analytics, International Journal of Informative & Futuristic Research, number = 10, pages = 3644–3654, volume = 2, year = 2015

22 W. D. Chambers (2014). *Computer simulation of dental professionals as a moral community. Medicine, Health Care and Philosophy* 17(3), 467–476.

23 Madeira, Lamont. Hoje a internet, amanhã os desafios da internet das coisas. 2011.

24 MAYUMI, Danielle. Computação nas nuvens – O futuro da internet. 2011.