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*—This article 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 concectam 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.**

Keywords— Big Data, Map Reduce, Hadoop, Swarm Robotics, Swarm Inteligence, Deep Learning, Machine Learning, Particle Swarm Optimization, Ant Colony Optimization, Fish School Search.

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 Luiza, 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*.

"Em 2010 empresas e usuários armazenaram mais de 13 exabytes de novos dados" [1].

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.

Para fazer frente a esse novo paradigma e extrair informações com eficiência uma nova abordagem algorítmica se faz necessário. As técnicas tradicionais de busca tradicionais 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 de 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.

Contudo muitos problemas de busca com técnicas heurísticas tradicionais nem sempre encontram a solução num tempo computacional aceitável. Felizmente a natureza tem contribuído como inspiração de soluções para vários problemas de buscas, muitos, de maneira eficaz e elegante. 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 bioinspirada 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.

2. Objetivos

Este artigo do tipo Survey tem como objetivo analisar a problemática do Big Data e seus paradigmas apresentados atualmente. Dentre esses podemos destacar o paradigma dos 5 V's; Volume de dados, Velocidade para acessar esses dados, Variedade de informações, Veracidade nos dados encontrados e Valor atribuído aos dados. Dentre as mais diversas ferramentas existentes destacamos algumas já consagradas como Map Reduce, da 'framework' Hadoop, e sua congênere desenvolvida pelo Google. Essas são as mais conhecidas dos "aventureiros" que estão desbravando o ainda pouco conhecido Big Data. Não será uma tarefa trivial inferir e desenvolver novas ferramentas, pois o Big Data tem assumido, nos mais recentes anos, proporções gigantescas como descrito na tabela 1

Tabela 1. 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}

3. Plano de execução do Survey

Para executarmos este Survey, foram primeiramente definidas as seguintes etapas:

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

A. Coleta dos dados
Para coletar os dados foram escolhidas seis (6) palavras chaves com relevância ao tema dentro do meio ambiente da pesquisa. Foram:

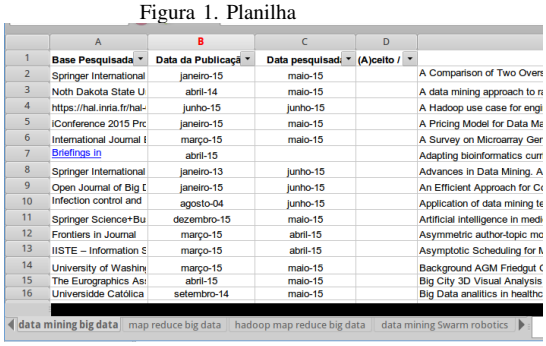
- "Data Mining and Swarm Inteligence"
- "Data Mining Big Data"
- "Data Mining Swarm Robotics"
- "Deep Learning"
- "Hadoop Map Reduce in Big Data"
- "Map Reduce Big Data"

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.

- 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;

Figura 1. Planilha



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.

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 2. 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 (Grande Rede Mundial) é apenas um dos muitos serviços que funcionam dentro da Internet.

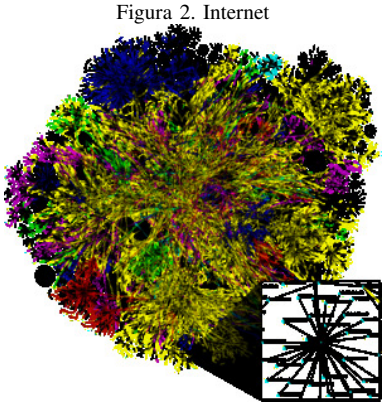
O acesso à Internet está condicionado a Provedores de Acesso (ISP), que se classificam em internacionais, que conectam países; nacionais (backbones criadas e mantidas por empresas especializads) ; regionais ligadas a outros ISPs, normalmente têm taxas de transmissão menores; e locais que se conectam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

5. Redes de Computadores

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 [3].

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

Figura 2. Internet



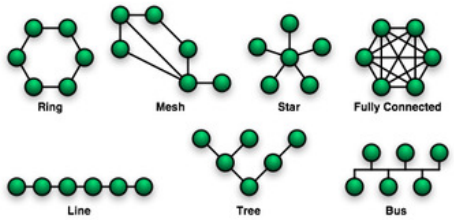
alguma catástrofe; e Segurança, que envolve proteção ao acesso de dados e proteção contra danos e perdas [3] [4].

5.1. Categorias de Rede

Quanto à Categoria, as redes são classificadas em:

a. **Redes pessoais (PAN)** – Conectam dispositivos em uma pequena área, aproximadamente $1m^2$;
b. **Redes locais (LAN)** – são privadas, podem englobar desde um escritório a um campus universitário. Alguns colocam a PAN nesta categoria de rede. As LANs são projetadas para compartilhamento de recursos computacionais;
c. **Redes geograficamente separadas (WAN)** - possibilitam transmissão de dados, imagens, áudio e vídeo por longas distâncias, áreas que podem compreender um país, um continente ou mesmo o mundo todo.
d. **Redes metropolitanas (MAN)** – rede de tamanho intermediário, entre LAN e WAN, abrangendo um distrito ou cidade. Normalmente usada para distribuir serviço de TV e Internet.

Figura 3. Topologia de rede



6. 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 [6] [6], 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

6.1. Particle 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 [7]. 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:

$$v_i(t+1) = v_i(t) + op_N best + op_L best op_N best = c_1 r_1 j(t)[Nbest - x_i(t)] op_L best = c_2 r_2 j[Lbest - x_i(t)] \tag{2}$$

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.

6.2. 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 [8] baseado no comportamento forrageiro de colônia de formigas. O comportamento forrageiro de diversas espécies [9] 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:

$$p(i,j) = \frac{[\tau(i,j)]^{\alpha}.[\eta(i,j)]^{\beta}}{\sum[\tau(i,j)]^{\alpha}.[\eta(i,j)]^{\beta}} \tag{3}$$

6.3. 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) [10].

Na busca FSS, cada peixe representa uma possível solução do problema ([10]). Em busca por enxame de partículas há o problema da degradação do exame, 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:

$$Bari(t) = \frac{\sum_{i=1}^{N} x_i(t) W_1(t)}{\sum_{i=1}^{N} x_i(t)} \tag{4}$$

7. Arquiteturas atuais

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 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)

8. Resultados esperados

Este Survey têm diversos propósitos preliminares e também conclusivos, para balizar nossa pesquisa, apontando o caminho mais promissor relacionado ao assunto escolhido, no nosso caso "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.

9. 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 [11] desde eletrodomésticos [12] a carros, sinais de trânsito, fazendo com que os *smartphones* seja a ferramenta para se ter acesso a isso tudo. [13] 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] 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

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

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

[4] E. SPECIALSKI (1999). *Gerência de redes de computadores e telecomunicações*

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

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

[7] 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

[8] 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

[9] 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

[10] 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

[11] Madeira, Lamont. Hoje a internet, amanhã os desafios da internet das coisas. 2011.

[12] MAYUMI, Danielle. Computação nas nuvens – O futuro da internet. 2011.

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