Fundação Educacional do Município de Assis
IMESA - Instituto Municipal de Ensino Superior de Assis

ANGELINA CASSIA DE PEDRI

DESMISTIFICANDO O MUNDO DO BIG DATA

Assis
2014

ANGELINA CASSIA DE PEDRI

DESMISTIFICANDO O MUNDO DO BIG DATA

Projeto de pesquisa apresentado ao Curso de Bacharelado em Ciência da Computação do Instituto Municipal de Ensino Superior de Assis – IMESA e da Fundação Educacional do Município de Assis – FEMA como requisito parcial a Pesquisa de Iniciação Cientifica – PIC.

Orientanda: Angelina Cassia De Pedri
Orientador: Prof. Dr. Alex Sandro Romeo de Souza Poletto
Co-Orientador: Fábio Girardi

Assis
2014

SUMÁRIO

1. Introdução 4
1.1. Problematização 4
1.2. Objetivos 5
1.3. Relevância/Justificativas 6
1.4. Estrutura do Trabalho 6
2. Banco de Dados 8
2.1. Introdução 8
2.2. Conceitos e Noções 9
2.3. Modelos de Bancos de Dados NoSQL 10
3. Big Data 12
3.1. Introdução 12
3.2. Conceitos e Noções 12
4. Data Mining 15
4.1. Introdução 15
4.2. Tarefas do Data Mining 17
4.2.1. Tarefas de Abordagem (Top Down) 17
4.2.2. Ferramentas para Data Mining 18
4.2.3. Técnicas para Data Mining 18
4.2.4.1. Rede Neural de Kohonen 19
4.2.4.2. Mecanismo de rede SOFM 20
4.2.4.3. Algoritmo da rede SOFM 20
5. Estudo de Caso (Data Mining): 22
6. Considerações Finais 25
7. Referencias Bibliográfica 26

1. INTRODUÇÃO

O termo Big Data está cada vez mais popular, porém, ainda não está bem claro o seu significado, a sua aplicabilidade e a sua finalidade. Para uma melhor compreensão, é essencial entender a definição dos 3V's: Volume + Variedade + Velocidade. O Volume representa a grande quantidade de dados gerados por sistemas corporativos, por mídias sociais, sensores e outros dispositivos; a Variedade representa os dados estruturados e não estruturados, obtidos do Twitter, Facebook, dentre outros, dados de empresas com grandes volumes de geração e movimentação de dados; a Velocidade, que representa a resposta quase que em tempo real para agir no próprio evento gerador das informações (TAURION 2013).

Big Data vem chamando atenção pela acelerada escala em que volumes cada vez maiores de dados são criados pela sociedade. No entanto, existem muitas dúvidas de como tangibilizar o conceito, ou seja, como sair do conceitual e criar soluções de negócio que mineralizem esta massa de dados, já que a cada dia são gerados dezenas de petabytes de dados, em uma escala real e não mais imaginária e futurista.

O Big Data nos permite a utilização de diversas tecnologias de gerenciamento dessa massiva geração de dados como, aplicação de Data Warehouse, Data Mart, Data Mining, Web Data Mining, BI (*Business Intelligence*) e Cloud Computing.

1.1. PROBLEMATIZAÇÃO

Atualmente essa quantidade massiva de dados vem de sistemas estruturados (que por sua vez são minoria) e não estruturados (maioria), gerados por emails, mídias sociais (Facebook, Twitter, Youtube, dentre outros), documentos eletrônicos, mensagens instantâneas, etc.

No entanto, as tecnologias atuais de gerenciamento de dados, como o modelo relacional proposto por Edgar F. Coode em 1969, não são mais adequadas para suportar os dados com a estrutura do Big Data. O modelo relacional é apropriado para acessar dados estruturados, gerados por sistemas internos das corporações, ou seja, não foi projetado

para dados não estruturados, já que não era esta a realidade da época, e nem para volumes na casa dos petabytes de dados como se tem nos dias de hoje.

Para tratar dados na escala de Volume, Variedade e Velocidade do Big Data, são necessários outros modelos mais apropriados como os sistemas gerenciadores de bancos de dados NoSQL, projetados para tratar imensos volumes de dados estruturados e não estruturados.

Existem diversos modelos como sistemas colunares como o Big Table, usado internamente pelo Google ou o modelo Key/value como DynamoDB da Amazon e o “document database” baseado no conceito proposto pelo Lotus Notes da IBM e aplicado em softwares como MongoDB.

Esta diversidade de alternativas demanda conhecimento e estudo sobre as tecnologias existentes com o intuito de aprimora-las para que seja possível tratar os dados não estruturados de uma maneira eficaz, considerando o Volume, Variedade e Velocidade com que esses dados são acessados.

1.2. OBJETIVOS

O objetivo geral é a desmistificação do conceito sobre Big Data que hoje é tido como um tema complexo para o público em geral e apresentar as vantagens das tecnologias relacionadas ao mesmo para tratamento de dados.

Sendo um tema que vem crescendo e atualizando-se a cada dia, terá também a apresentação e exemplificação de uma das tecnologias para mineralização de dados que será o Data Mining.

Além disso, pretende-se:

- Ampliar os conhecimentos sobre Big Data;
- Compreensão e divulgação dos conceitos envolvidos na mineralização de dados;
- Desenvolver uma explanação sobre o conceito de Big Data;
- Desenvolver uma analise sobre uma das formas de aplicação da tecnologia para mineralização de dados que seria Data Mining;

- Apontar vantagens e possibilidades para utilização desta tecnologia para empresas de pequeno a grande porte.

1.3. RELEVÂNCIA / JUSTIFICATIVAS

O intuito desta pesquisa é sintetizar os principais conceitos relacionados ao Big Data e tecnologias que o abrangem. Espera-se que este trabalho apresente como resultado uma base introdutória para conhecimento do tema citado acima, podendo ser utilizado como um guia, para familiarização com o assunto.

Tendo em vista o interesse e as dificuldades para compreensão do Big Data, a existência de um trabalho que aborde e analise o tema como este pode ser relevante e justo pelo fato de auxiliar na introdução de profissionais ou leigos ao tema ampliando sua grade de conhecimento voltada para o armazenamento de dados.

Este segmento no Brasil é novidade, tendo em vista que o mesmo vem sendo estudado e divulgado há cinco anos no país e utilizando este projeto como guia introdutório o profissional da área de TI terá maior conhecimento sobre Big Data, mesmo que seja um tanto superficial, olhando pela imensa quantidade de tecnologias para aplicação e seus conceitos.

1.4. ESTRUTURA DO TRABALHO

Este trabalho está dividido em sete capítulos sendo essa Introdução o primeiro capítulo.

No segundo capítulo serão apresentados os conceitos relacionados ao SGDB (Sistema Gerenciador de Banco de Dados) NoSQL.

O terceiro capítulo contemplará explanações sobre Big Data e os fatores que abrangem esta tecnologia.

No quarto capítulo serão abordados os conceitos sobre Data Mining.

No quinto capítulo será apresentado o estudo de caso.

E para finalizar a contextualização, o sexto capítulo tratara das considerações finais.

Por ultimo, no sétimo capitulo, serão relacionadas às referências bibliográficas.

2. BANCO DE DADOS NoSQL

2.1. INTRODUÇÃO

Até o presente momento, a bibliografia consultada não contemplou diretamente nenhum dos objetivos que se espera atingir ao término da pesquisa. Os textos abordados, entretanto, servem como apoio para um ingresso no conceito de Big Data.

Pouco a pouco, passa-se agora a destacar a contribuição que cada artigo conferiu ao projeto até esse ponto. O artigo de TAURION (2013) discorre sobre a abordagem do tema Big Data e sua importância na atualidade quanto à imensa geração de dados estruturados e não estruturados a cada dia e uma breve abordagem sobre a tecnologia para minerar dados, Data Mining, mostrando como sua aplicação em um servidor que armazena dados a nível de Big Data se faz possível à sumarização das informações diminuindo o tempo de processamento das mesmas.

Já a apresentação de BARTH (2012) tem início com uma rápida abordagem sobre a mineralização de dados utilizando a tecnologia Web Data Mining. O autor centra o seu esforço nessa última categoria, o Web Data Mining tem como objetivo mineralizar e aperfeiçoar o gerenciamento de operações corporativas diárias.

Ainda de acordo com BARTH, o Web Data Mining é capaz de sumarizar os dados a nível de Big Data e extrair do banco de dados apenas a informação sumarizada para a consulta feita no momento.

Os autores GOLDMAN et al (2011) em pesquisa realizada no artigo “Apache Hadoop: conceitos teóricos e práticos, evolução e novas possibilidades” tratam a mineração de dados das redes sociais utilizando o framework Apache Hadoop para o mesmo mostrando a eficácia obtida por esta plataforma onde pode ser constatada ao verificar a quantidade de importantes empresas, de diferentes ramos, que estão usando Hadoop para fins educacionais ou de produção, como exemplo a Yahoo!.

2.2. CONCEITOS E NOÇÕES

Os modelos de Bancos de Dados determinados como NoSQL são diferentes sistemas de armazenamento que vieram para suprir necessidades nas demandas em que os bancos de dados tradicionais (relacionais) não são muito eficientes.

Muitas bases de banco de dados não relacionais apresentam características bem interessantes como alta performance, escalabilidade, replicação, suporte à dados estruturados e sub colunas.

O modelo NoSQL ou NoRel surgiu da necessidade de um desempenho superior e de uma alta escalabilidade, pois os atuais bancos de dados relacionais são muito restritos a este sistema de armazenamento, sendo necessário a distribuição vertical de servidores, ou seja, quanto mais dados, mais memória e mais disco um servidor precisa.

Este modelo tem uma grande facilidade na distribuição horizontal, ou seja, mais dados, mais servidores, não necessariamente de alta performance, sendo essa forma de utilização muito mais eficiente e econômica.

Portanto, os bancos de dados NoSQL possuem toda a informação necessária agrupada no mesmo registro, ou seja, em vez de ter o relacionamento entre varias tabelas para formar uma informação, ela estará em sua totalidade no mesmo registro.

Além das vantagens citadas acima, os bancos de dados NoSQL são muito tolerantes a erros.

Um exemplo de eficiência e usabilidade dos bancos NoSQL é a utilização do mesmo para atender os usuários de empresas que utilizam vários Data Centers, localizados em diversas partes do pais ou do mundo.

Com isso, uma série de questões sobre disponibilidade e performance é levantada ao construir os sistemas.

A distribuição destes sistemas combinada com o hardware não tão caro e de alta performance, impõe ao sistema a necessidade de ser robusto o suficiente para tolerar

falhas constantes e imprevisíveis, seja de hardware, seja da infraestrutura do lugar onde o Data Center se encontra.

Pensando nessas questões, bem como nas necessidades internas ou dos clientes, foi surgindo uma grande quantidade de bancos de dados não relacionais de trabalham de diferentes maneiras cada um com seu modelo de armazenamento de dados, que serão citados no capitulo a seguir.

2.3 MODELOS DE BANCO DE DADOS NOSQL

Antes da citação dos modelos de bancos de dados NoSQL mais utilizados, vale ressaltar as características que definem um Banco de Dados Relacional e não relacional.

Os Bancos de Dados Relacionais baseiam-se no armazenamento dos dados em tabelas, conceito de entidade e relacionamento.

Os dados são separados de forma única, tentando diminuir ao máximo a redundância, já que as informações são criadas pelo conjunto dos dados, onde são as relações entre as tabelas que executam esta tarefa.

Já os bancos de dados não relacionais possuem uma solução alternativa para os bancos de dados relacionais como alta escalabilidade e desempenho, conforme citado nos capítulos anteriores.

Os Bancos de Dados NoSQL são subdivididos pelo seu núcleo, ou seja, a maneira de como ele trabalha com os dados como armazenamento e organização.

A seguir, serão apresentados alguns dos modelos NoSQL mais conhecidos, focados na funcionalidade e aplicabilidade.

Key/Value Store (Armazéns Chave-Valor)

Esse é o tipo de banco de dados mais simples, já que o conceito dele é uma chave e um valor para essa chave, ou seja, são sistemas distribuídos nessa categoria, também conhecidos como tabelas de hash distribuídas, armazenam objetos indexados por chaves, e possibilitam a busca por esses objetos a partir de suas chaves.

Este modelo suporta uma grande carga de dados, sendo assim possuem maior escalabilidade.

Alguns bancos de dados baseados em Key/Value Store mais conhecidos são: Berkley DB, Tokyo, Cabinet, Project, Voldermort, MemcacheDB, SimpleBD e Wide.

Columns Store (Armazenamento orientado a Colunas)

Fortemente inspirados pelo BigTable do Google, eles suportam várias linhas e colunas, além de permitir subcolunas.

Além do BigTable outros bancos que usam essa tecnologia são HBase (Apache), HiperTable e Cassandra (Apache).

Document Store (Armazenamento orientado a Documentos)

Baseado em documentos XML ou JSON, podem ser localizados pelo seu ID único ou por qualquer registro que tenha no documento.

Portanto, os documentos dos bancos de dados dessa categoria, são conjuntos de atributos e valores, onde um atributo pode ser multivalorado.

Em geral, os bancos de dados orientados a documento não possuem esquema, ou seja, os documentos armazenados não precisam possuir estrutura em comum. Essa característica faz com que seja uma boa opção para o armazenamento de dados semiestruturados.

Alguns bancos de dados que utilizam este recurso são CouchDB (Apache), MongoDB, Riak e RavenDB.

Graph Store (Armazenamento orientado a Grafos)

Diferentemente de outros tipos de Bancos de Dados NoSQL, esse está diretamente ligado a um modelo de dados estabelecido, o modelo de grafos.

A ideia desse modelo é representar os dados ou o esquema dos dados como grafos dirigidos, ou como estruturas que generalizem a noção de grafos, contendo três componentes básicos: os nós (são os vértices do grafo), os relacionamentos (são as arestas) e as propriedades (ou atributos) dos nós.

Neste caso, o banco de dados pode ser visto como um multigrafo rotulado e direcionado, onde cada par de nós pode ser conectado por mais de uma aresta, ou seja, guardam objetos e não registros como os outros tipos de NoSQL, contemplando a busca desses itens é pela navegação dos objetos.

Os bancos de dados que utilizam o mesmo conceito são Neo4J, InfoGrid, HyperGraphDB e BigData.

3. BIG DATA

3.1 INTRODUÇÃO

As explanações a seguir contemplam os conceitos básicos de Big Data, mostrando onde pode ser aplicado e o motivo de sua utilização.

O termo Big Data vem ganhando grande ênfase dentro das pequenas, médias e grandes empresas, colocando como proposta a melhoria na organização da vasta massa de dados conferida no dia-a-dia ou então para utilizar estes dados a favor de pequenas empresas, .

Como contribuição para esta ideia tem-se os artigos de Gasparotto,Henrique M. e Ianni, Vinicius, onde abordam de maneira sucinta o tema de armazenamento de dados em grande escala, com a utilização do Hadoop e o MapReduce.

3.2 CONCEITOS E NOÇÕES

Big Data, é o termo utilizado atualmente para nomear a grande quantidade de dados armazenados em servidores vindos de diversas fontes de dados como mídias sociais (Twitter, Facebook, e-mails), sensores, e assim por diante.

Estes dados que antes não tinham valor, hoje com a solução NoSQL é possível tirar grande aproveito dos mesmos.

Esta solução vem sendo chamada de Big Data, que para relembrar este termo o mesmo é definido como 3Vs, volume+variedade+velocidade, onde:

Volume: representa as informações geradas pelos sistemas transicionais somadas aos dados gerados pelos sensores, câmeras, mídias sociais, via smartphones, tablets, entre outros meios de comunicação utilizados.

Variedade: define a forma que estes dados são apresentados, já que podem ser estruturados, semiestruturados, não estruturados tais como fotos, e-mails, logs, posts, e demais.

Velocidade: na maioria das vezes as respostas aos eventos precisam ser praticamente em tempo real e com isso, tratando um volume massivo de dados na casa de terabytes e zetabytes.

Para esses dados “desordenados” existem os Bancos de Dados NoSQL que auxiliam na organização e armazenamento, ou seja, bancos de dados não relacionais que tem a capacidade de armazenar os dados estruturados, semi estruturados e não estruturados em melhor escala, utilizando o modelo de grafos e chave/valor (key value), conforme apresentado no Capítulo 2.

O conjunto das informações citadas mais a solução NoSQL é que formam o Big Data, um modelo de bando de dados não relacional.

Hoje em dia, é possível observar os benefícios de negócios tangíveis e pragmáticos com o uso do Big Data, seja para aumentar a taxa de conversão para reservas, diminuir custos de operação, impulsionar receita ou elevar a satisfação do consumidor.

Por exemplo, para uma empresa área, com a tecnologia em tempo real e o armazenamento de grande quantidade de dados que o Big Data possibilita, é possível ter o perfil de cada passageiro, sabendo o destino mais procurado, tipo de acento, época de viagens, e demais informações, utilizando as características de cada um no momento da compra das passagens.

Assim esta empresa pode encaminhar pacotes promocionais e demais vantagens que são direcionadas exatamente para o perfil do cliente selecionado. Assim estará conquistando mais clientes de acordo com as escolhas personalizadas e aumentando também os lucros.

4. DATA MINING

4.1 INTRODUÇÃO

Data Mining, consiste em um processo analítico projetado para explorar grandes quantidades de dados, geralmente armazenadas em um Data Warehouse, na busca de padrões consistentes e relacionamentos sistemáticos entre variáveis e validá-los aplicando os padrões a novos subconjuntos de dados.

Este processo consiste basicamente em 3 etapas: exploração, construção de modelo ou definição do padrão e validação.

Os requisitos para utilização do Data Mining é obter uma argumentação ativa, onde em vez do usuário definir o problema, primeiramente selecionar os dados e as ferramentas para analisar destes dados.

As ferramentas do Data Mining pesquisam automaticamente os mesmos a procura de anomalias e possíveis relacionamentos, identificando assim problemas que não tinham sido identificados pelo usuário.

A Figura 1 apresenta os passos do Data Mining que auxiliam na identificação dos problemas diários.

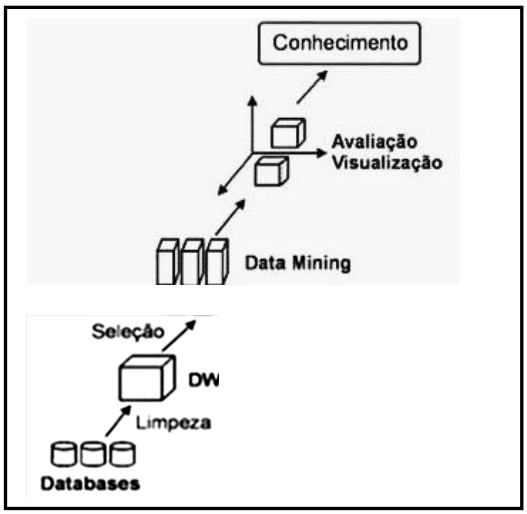

Figura 1. Passos do processo de Data Mining

Estes passos também analisam os dados, descobrem problemas ou oportunidades escondidas nos relacionamentos dos dados, diagnosticam o comportamento dos negócios, requerendo a mínima intervenção do usuário.

As ferramentas são baseadas em algoritmos que formam a construção de blocos de inteligência artificial, que facilitam e auxiliam o trabalho dos analistas de negócio das empresas.

Abaixo na Figura 2, será exemplificado o esquema para analise dos problemas levantados pelo usuário, que são verificados pelo analista.

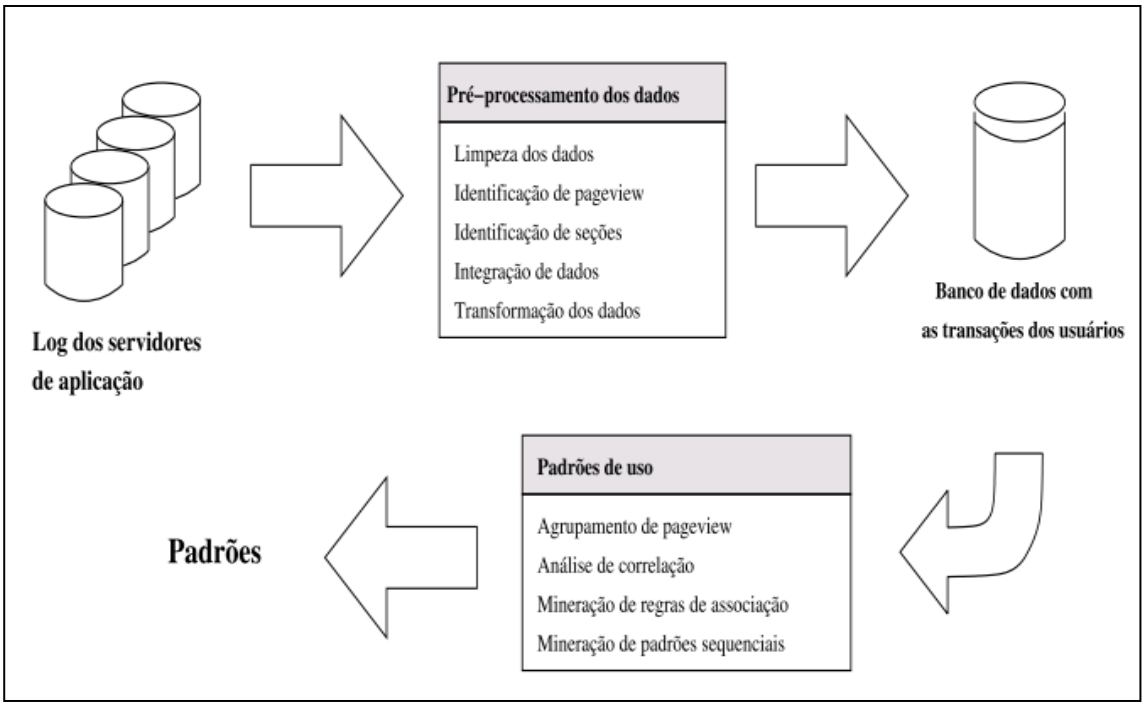

Figura 2. Esquema para Análise do Problema

A seguir são apresentadas algumas tarefas, ferramentas de analise de dados e técnicas utilizada no Data Mining para detectar problemas no processamento.

4.2 TAREFAS DO DATA MINING

As tarefas do DM são classificadas de acordo com a abordagem que se deseja seguir.

Dependendo da abordagem e regra de negócio estabelecida nem todas as tarefas são obrigatórias, e em muitos casos muitas podem ser agrupadas constituindo uma só.

A seguir serão mostradas as tarefas de acordo com suas abordagens.

4.2.1 TAREFAS DE ABORDAGEM (TOP DOWN)

Neste tipo de tarefa tem-se a Estimação, Predição e Classificação. A Estimação é o processo de predizer algum valor baseado em um padrão já conhecido. Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais. A Predição prevê um comportamento futuro baseado em vários valores. A Classificação identifica algum valor para uma variável categórica.

Na tarefa TOP DOWN, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de classificar um novo registro.

Os dados podem ser associados à classe pelo processo de discriminação, onde o resultado obtido provém de um valor atribuído a um registro em função de um ou mais atributos do mesmo, ou por caracterização, pela sumarização de um atributo de estudo por uma característica de um ou mais atributos.

4.2.3 FERRAMENTAS PARA DATA MINING

Ferramentas disponíveis para auxiliar no processo do Data Mining:

Ferramenta	Fornecedor	Tarefas
WEKA	University of Waikato	Classificação, Regressão e Regras de Associação.
Intelligent Miner	IBM Corp.	Classificação, Regras de Associação, Clusterização e Sumarização.
Oracle Data Miner	Oracle	Classificação, Regressão, Associação, Clusterização e Mineração de Textos.
SAS Enterprise Miner Suite	SAS Inc.	Classificação, Regras de Associação, Regressão e Sumarização.
Clementine	SPSS Inc.	Classificação, Regras de Associação, Clusterização, Sequência e Detecção de Desvios.
Darwin	Thinking Machines	Classificação.
Business Objects	Business Objects	Classificação, Regras de Associação, Clusterização e Sumarização.
Microsoft Data Analyser	Microsoft Corp.	Classificação e Clusterização.
MineSet	Silicon Graphics Inc.	Classificação, Regras de Associação, Análise Estatística.
DBMiner	DBMiner Technology Inc.	Classificação, Regras de Associação e Clusterização.
Gemanics Expression	Gemanics Developer	Análise de Sequências.
SAS Text Miner	SAS Inc.	Mineração de Textos.

4.2.4 TECNICAS PARA DATA MINIG

Técnicas para o levantamento dos problemas:

Tarefa	Técnica
Estimação	Regressão Linear, Múltipla, não Linear, Logística, Poisson.
Predição	Regressão Linear, Múltipla, não Linear, Logística, Poisson.
Classificação	Árvore de Decisão, Classificação Baeysiana, Rede Neural, Classificação por Regras, Análise de Vizinhança, Algoritmos Genéticos, Lógica Fuzzy.

4.2.4.1 REDE NEURAL DE KOHONEN

De acordo com Lippmann (1987), as redes *SOFM*, conhecido por *Self-Organising Feature Maps* (Mapas de Características Auto-Organizados), desenvolvido por Teuvo Kohonen na década de 80, são do tipo *feedforward* (acíclica – alimentação adiante) com aprendizagem não supervisionada.

Essas redes usam simples unidades de processamento adaptativas para receber sinais de um ambiente externo. Esses sinais compõem-se de medidas ou dados, como frequência ou situação.

Baseado em Haykin (2001), pode-se afirmar que esse modelo é considerado uma classe especial de *mapas neurais*, baseada na aprendizagem competitiva.

As redes neurais artificiais são consideradas ferramentas de grande capacidade de aprendizagem, via treinamento, em diversas áreas do conhecimento. Elas são desenvolvidas para reconhecer e classificar padrões informacionais.

Nelas, as unidades de processamento da camada de saída competem entre si para serem ativadas ou disparadas, por meio de uma função de ativação. O resultado é que apenas uma unidade de processamento de saída sairá vencedora. Trata-se de um tipo de competição induzido por conexões laterais inibitórias (negativas) entre as unidades de processamento.

Já Kohonen (1987) observa que o esquema básico de seu modelo *SOFM* faz com que unidades de processamento da camada intermediária disputem entre si a representação de informação fornecida pelas unidades de processamento de entrada, a fim de se tornarem a unidade vencedora. Havendo uma unidade de processamento vencedora, esta é reajustada para responder ainda melhor ao estímulo recebido. Este modelo não somente ajusta o vencedor, mas também as unidades de processamento vizinhas à unidade vencedora.

O modelo *SOFM* é baseado no mapa topológico presente no córtex cerebral(LIPPMAN, 1987), que mantém uma ordem entre suas unidades de processamento (neurônios). Ou

seja, áreas responsáveis por funções especificas onde possuem subáreas que mapeiam de forma ordenada as entradas das informações.

4.2.4.2 MECANISMO DE REDE *SOFM*

Segundo Kohonen (1989), o mecanismo da rede *SOFM* funciona da seguinte forma: os pesos sinápticos iniciam contendo valores aleatoriamente baixos; um estímulo de entrada x é provido para a rede, sem que se especifique a saída desejada.

O fato de não se especificar uma saída desejada caracteriza a rede como não supervisionada ou auto-organizada. O estímulo de entrada x é descrito como um vetor x=(x1, x2, x3, ..., xn). Quando o estímulo de entrada é introduzido na rede, ele é disputado pelas unidades de processamento de saída $\vec{y}$. A unidade de saída que melhor responder ao estímulo de entrada, será definida como unidade de saída vencedora.

A rede é considerada treinada depois que todo o conjunto de estímulos de entrada do vetor x de treinamento tiver sido apresentado à rede e satisfeita os critérios de aprendizagem.

4.2.4.3 ALGORITMO DA REDE *SOFM*

Conforme Lippmann (1987), o algoritmo de aprendizagem da rede *SOFM* de Kohonen também definido como algoritmo de aprendizado competitivo, opera numa sequência de seis passos durante o treinamento, os quais podem ser assim descritos:

Primeiro passo: corresponde à iniciação dos pesos das conexões w=(x11, x21,..., xnj), com valores aleatórios próximos de zero. Os pesos aleatórios, em certos casos, levam a um processo muito lento de convergência. Dessa forma, pode ocorrer a iniciação da rede pela distribuição dos pesos uniformemente pelo espaço de pesos para acelerar o processo.

Também no primeiro passo é iniciado o raio da vizinhança da unidade de processamento j, chamado de Vj(t), para um valor de raio abrangente.

Segundo passo: realiza-se a apresentação dos estímulos de entrada x=(x1(t), x2(t),..., xn(t)), sendo x o vetor das unidades de processamento que estão no tempo t.

Terceiro passo: efetua-se o cálculo da distância euclidiana entre o padrão do estímulo de entrada e cada unidade de processamento da rede. Ou seja, toma-se uma unidade de processamento com índice 2, os estímulos de entrada apresentados por X=[1,3,2,4] e os respectivos pesos, apresentados por W=[0,1,0,1]. Segundo Haykin (2000, p. 51), a distância euclidiana entre o par de vetores, padrão do estímulo de entrada e pesos é definida por:

$$d(X_i, W_j) = ||X_i - W_j||$$

$$= \left[\sum_{k=1}^{n} (X_{ik} - W_{jk})^2 \right]^{1/2}$$

Quarto passo: seleciona-se a unidade de processamento que apresentou a saída com menor distância euclidiana. Esta unidade de processamento é designada como a unidade vencedora.

Quinto passo: atualizam-se os pesos da unidade de processamento juntamente com os pesos de todas as unidades de processamento que estão dentro da vizinhança.

Sexto passo: se o número de iterações realizadas até o momento atingiu o número especificado no início, encerra-se a fase do treinamento. Caso contrário, repete-se o processo a partir do segundo passo.

5. ESTUDO DE CASO (DATA MINING): Analise da Proposição sob uma Base de Dados Hipotética baseada em Tese

Para o estudo de caso foi utilizada uma massa de dados hipotética, representando uma população de materiais a serem classificados através de *clusters*, levando em consideração altura e peso do material.

Os dados foram preparados e processados em uma Rede Neuronal utilizando o software MathLab.

Observou-se, inicialmente, a intervenção humana durante a fase de preparação dos dados, limpeza e carga, onde os dados foram carregados em uma matriz.

A Figura 3 ilustra, por meio de um diagrama de dispersão, as entradas na matriz:

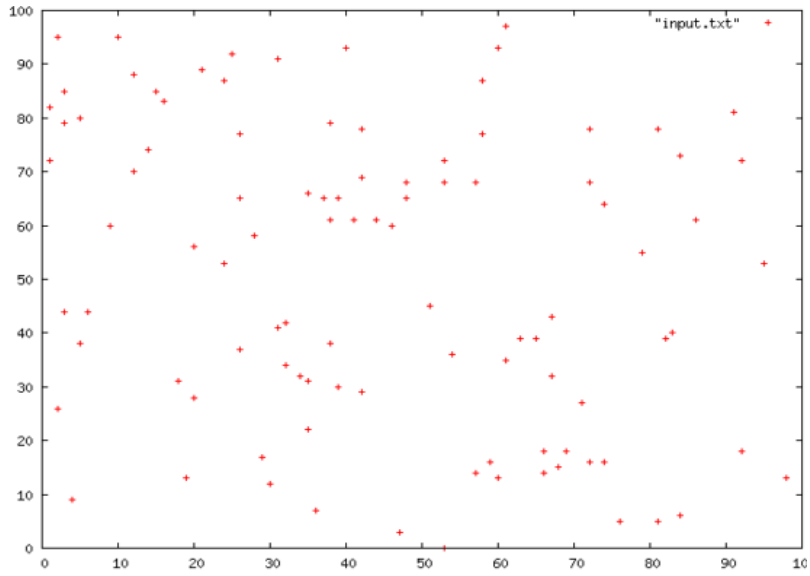

Figura 3. Plotagem dos pontos antes da classificação

Após carregar as informações, o usuário criou a rede neuronal, conhecendo as mesmas ocultadas pelo volume da massa de dados.

Para criar a rede, foi necessário a escolha do algoritmo, a definição de parâmetro, e apontar para a base de dados preparada. Após isso foi realizado o treinamento, que

ajustou os pesos entre as sinapses de forma a carregar o conhecimento necessário para, neste caso de *clusterização*, efetuar a classificação.

No caso desta rede neuronal para classificação em *clusters*, foi utilizado o algoritmo Kohonen (rede competitiva), onde o treinamento é não supervisionado – é feito sem que exista uma massa de dados já classificada (sem intervenção humana), sendo os pesos ajustados através de interações temporais e novos ajustes, de forma a estabilizar em um espaço de tempo.

Nesse algoritmo, cada neurônio aprende a responder maximamente a diferentes valores de entrada, ou seja, numa massa de dados plotadas em um gráfico de dispersão, procura centros que diminuam a distancia dos pontos dentro da “classe” que o neurônio responde, e aumente a distância entre as classes (excitação central – inibição lateral).

O MathLab apresenta graficamente (Figura 4) as interações e posicionamento do “centro” do cluster conforme as épocas, e com isso, o usuário pode acompanhar o ajuste.

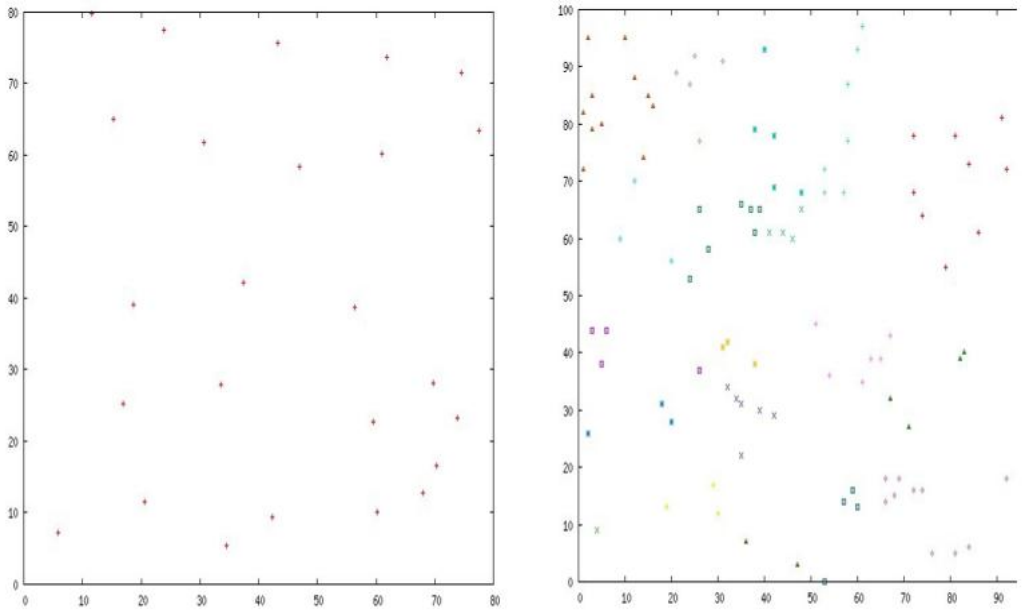

Figura 4. A esquerda (centros de gravidades das classes). A direita (os pontos classificados)

Porém, o MathLab não fornece uma interface para que, no meio da execução da rede, sejam alterados pontos que interfiram na formação ou interferências que não foram identificadas na limpeza dos dados.

A Figura 5 ilustra alterações a partir da intervenção do analista humano no processo de condução e interpretação da mineração.

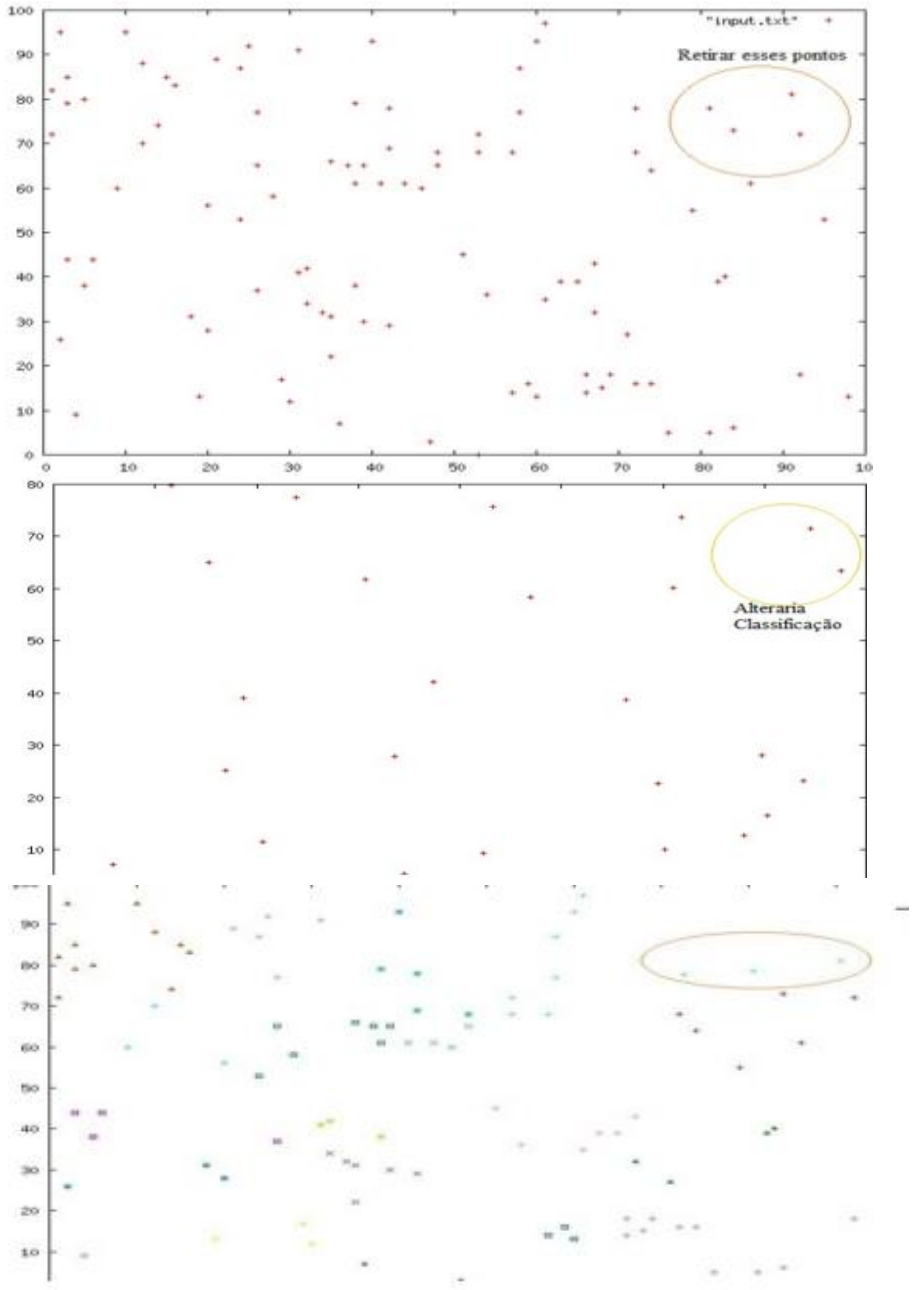

Figura 5. Influências da intervenção humana sobre os resultados

6. CONSIDERAÇÕES FINAIS

Pode-se observar no estudo de caso apresentado, que a eficiência algorítmica dos processos de Data Mining a frente de volumosos e complexos conjuntos de dados como no caso do Big Data, juntamente com a capacidade de analise cognitiva dos especialistas formam uma poderosa ferramenta de analise de dados, auxiliando na verificação de problemas cotidianos, atendendo as expectativas organizacionais.

A contribuição da analise por especialista utilizada neste estudo, gerou maior dinamismo no sistema, trazendo resultados finais mais satisfatórios, onde trazem a realidade do dia-a-dia.

A elaboração de uma interface gráfica permitirá a percepção visual e em número suficientemente diminuído de elementos para que o analista possa utilizar sua memória de trabalho, como foi exemplificado no estudo citado.

REFERÊNCIAS BIBLIOGRÁFICAS

TAURION, Cezar. **Coletânea de posts publicados no Blog developerWorks em2012developerWorks Brasil**. Disponível em: <http://www.ibm.com/developerworks/blogs/page/ctaurion >. Acesso em: 07 outubro. 2013.

BARTH, J. Fabrício. **Uma Introdução á Mineração de Informações na era do Big Data**. 2012. 75. Tipo de trabalho (Titulação) - VAGASTecnologia e Faculdades BandTec.

SBPJor – Associação Brasileira de Pesquisadores em Jornalismo, IX, 2011, Rio de Janeiro. **Jornalismo Computacional em função da Era do Big Data**: 2011.12.

GOLDMAN, A., KON, F., JUNIOR, F. P., POLATO, I, PEREIRA, R. F**.. Apache Hadoop: conceitos teóricos e práticos, evolução e novas possibilidades**.

KIMBALL, R. **The Data Warehouse Toolkit**: Guia Completo para modelagem dimensional. Tradução da segunda edição. Rio de Janeiro: Campus Ltda, 2002.

BARASUOL, Érion Ricardo. **MongoDB uma base de dados orientada a documentos que utiliza orientação a objetos**. 2012. 101 f. Trabalho de Conclusão de Curso - Instituto Municipal de Ensino Superior de Assis, 2012.

DOURADO, Joana. **Semantix – Treinamentos.** Disponível em: <http:// www.semantix.com.br>. Acesso em: 05 dezembro. 2013.

CAPETTA, M. Leonardo. **Consulta a banco de dados em linguagem natural**. Omnia Exatas, v.4, n.1, p.72-80, 2011.

GOMES, M. Heitor., HAUTH, G. Luiz., CARVALHO, R. Deborah. **Mineração de dados temporal: Descoberta de Regras de Causa e Efeito**. Faculdade de Ciências Exatas e Tecnologia (FACET) – Universidade Tuiuti do Paraná – Curitiba – PR – Brasil.

BARTH, J. Fabrício. **Uma Introdução a Mineração de Informações na era do Big Data**. VAGAS Tecnologia e Faculdades BandTec.

NASCIMENTO, O., J., Rafael. **Mineração e Análise de Dados em SQL**. Disponível em: <http://www.devmedia.com.br/mineracao-e-analise-de-dados-em-sql/29337>. Acesso em: 04 de Agosto de 2014.

BIRGNOLI, T. Juliano, JUNIOR, S. Egon, MIGUEZ, B. Viviane, SANTOS, Neri, SPANHOL, Fernando. **A Intervenção Humana na Qualificação de Processos de Data Mining: Estudo de Caso em uma Base de Dados Hipotética.** Universidade Federal de Santa Catarina.

MAYER-SCHÖNBERGER, Viktor., CUKIER, Kenneth. **BIG DATA, Como Extrair Volume, Variedade, Velocidade e Valor da Avalanche Cotidiana**. Editora Elsevier Ltda, 2013.

KOHONEN, Teuvo**. An Introduction to Neural Computing**. Finland: Helsinki University of Tecnology, pp. 3-16, 1988.

KOHONEN, Teuvo**. Self-Organization and Associative Memory**. 2ª Edição. USA: Spring-Verlag, 1989.

KOHONEN, Teuvo. **The self-organizing map**. Proceedings of the Institute of Electrical and Electronics Engineers, vol.78, pp. 1464-1480, 1990.

LIPPMANN, Richard. **An Introduction of Computing with Neural Nets**. IEEE Computer Society, v.3, nº 4, pp. 4-22, abr.1987.