Evento: XXII Jornada de Pesquisa

UM MODELO HÍBRIDO BASEADO EM CLASSIFICAÇÃO E CLUSTERIZAÇÃO PARA A EXTRAÇÃO DE INFORMAÇÃO A PARTIR DE GRANDES BASES DE DADOS[1]

A HYBRID MODEL BASED ON CLASSIFICATION AND CLUSTERIZATION FOR THE EXTRACTION OF INFORMATION FROM LARGE DATABASES

Patricia Mariotto Mozzaquatro Chicon[2], Fabricia Carneiro Roos Frantz[3]

[1] Pesquisa desenvolvida no Departamento de Ciências Exatas e Engenharias (DCEEng), pertencente ao Grupo de Pesquisa Applied Computing Research Group (GCA)
[2] Aluna do Curso de Doutorado em Modelagem Matemática, email:patriciamozzaquatro@gmail.com
[3] Professora Doutora do Departamento de Ciências Exatas e Engenharias, email: frfrantz@unijui.edu.br

RESUMO

Com o surgimento da big data, as organizações empresariais estão adotando uma nova postura para lidar com o grande volume de dados, sendo eles estruturados ou não estruturados. Neste contexto, torna-se necessário a utilização de mecanismos computacionais que permitam a interação com estes dados de forma inteligente e rápida, assim surgindo a ciência de dados. O objetivo deste artigo é propor um modelo híbrido baseado em classificação e clusterização para a extração de informação a partir de grandes bases de dados. Os métodos j48 e k-means, utilizados na mineração de dados, irão respectivamente, classificar e agrupar as informações similares de acordo com a necessidade organizacional a fim de gerar uma tomada de decisão eficaz.

ABSTRACT

With the emergence of the big data, business organizations are taking a new stance to deal with the large amount of data, whether structured or unstructured. In this context, it becomes necessary to use computational mechanisms that allow the interaction with this data in an intelligent and fast way, thus arising the science of data. The objective of this article is to propose A hybrid model based on classification and clustering for the extraction of information from large databases. The j48 and k-means methods, used in data mining, will respectively classify and group similar information according to organizational need to generate effective decision making.

Palavras – Chave: Modelo Híbrido. Mineração de Dados. Classificação. Clusterização.
Keywords: Hybrid Model. Data Mining. Classification. Clustering.

1. INTRODUÇÃO

O volume de dados produzidos, armazenados ou transmitidos no mundo tem crescido exponencialmente nos últimos anos. O surgimento dessa enorme quantidade de dados constitui o que se chama atualmente de big data, sendo considerado o quarto paradigma da ciência, (PALETA, 2014). Conforme Gantz e Reinsel (2011), big data é um corte horizontal do universo digital e pode incluir dados transacionais, dados armazenadas, meta dados e outros dados que

residem em arquivos muito grandes.

As organizações empresariais estão adotando uma nova postura para interagir com o grande volume e variedade de dados, provenientes de fontes como dispositivos móveis, mídias sociais, dentre outras. Estes dados podem estar tanto estruturados, quanto não-estruturados, produzidos diariamente, de modo a subsidiar melhores decisões estratégicas. Nesse cenário, a computação tem se tornado uma importante ferramenta de trabalho para pesquisadores de diversas áreas.

Como resultado destas transformações no meio empresarial, e como forma de responder às demandas existentes, observa-se a expansão de uma área de estudo, interdisciplinar e intensivamente computacional: a ciência de dados (Data Science) (CURTY, CERVANTE, 2016). Esta ciência visa estudar os dados, seu processo de captura, transformação, geração e, posteriormente, análise.

O termo Data Science, criado em 2010, corresponde ao que em 1970 chamava-se de Decision Support Systems (DSS), nos anos 80 aos Executive Information Systems (EIS), nos anos 90 aos Online Analytical Processing (OLAP), e nos anos de 2000 ao Business Intelligence (BI) (DAVENPORT, 2014). De acordo com Grus (2016), Data Science é o termo atual para a ciência que analisa dados, combinando a estatística com machine learning/data mining e tecnologias de base de dados, para responder ao desafio que o big data apresenta. Dessa forma, como definido por Curty e Serafim (2016), Data Science é a ciência que estuda metodologias e técnicas a fim de inferir informações a partir de grandes bases de dados (ou big data).

Conforme estudos de Curty e Serafim (2016), a expectativa é de que em 2017 existam cerca de cinco (5) bilhões de telefones celulares conectados, os quais serão responsáveis por produzir um fluxo gigantesco e constante de informação digital. A estimativa afirma que cerca de 90% de todos os dados armazenados em todo o mundo foram produzidos somente nos dois (2) últimos anos e seus rastros continuam se multiplicando a cada ano que passa.

Nesta perspectiva, os autores Goldschmidt et.al (2015) relatam os aspectos relacionados as dificuldades de mineração e atribuição de sentido automático ao tratar grandes quantidades de dados textuais em linguagem natural.

Analisando estes aspectos e os trabalhos em desenvolvimento na área, acredita-se que por meio da integração das técnicas de mineração de dados classificação e clusterização, será possível extrair e filtrar de grandes bases de dados conhecimento inédito, útil e relevante existente em conjuntos de dados. Portanto, este artigo propõe um modelo híbrido baseado em classificação e clusterização para a extração de informação a partir de grandes bases de dados.

A necessidade de desenvolver novas tecnologias de processamento para big data surgiu em vários domínios devido ao aumento do poder computacional, capacidade de armazenamento e aumento na produção de conteúdo digital. Executar tarefas de Mineração de Dados no contexto da big data não é uma tarefa trivial. Assim justifica-se a pesquisa aqui apresentada.

Este artigo está organizado em quatro seções. A Seção 2 apresenta um estudo teórico sobre: Ciência de Dados, a partir de alguns de seus conceitos, além de processos, tarefas e técnicas; descreve a descoberta de conhecimento em grandes bases de dados, abordando a tarefa de mineração de dados juntamente com as técnicas de classificação, descrevendo o algoritmo J48 e a técnica de clusterização, também conceituando o algoritmo k-means. Apresenta um estudo sobre as pesquisas desenvolvidas na área. A Seção 3 descreve a metodologia abordando as etapas da pesquisa como também a proposta de desenvolvimento do modelo híbrido. A Seção 4 contém a conclusão do artigo, e, finalmente são descritas as referências.

2. REFERENCIAL TEÓRICO

Nesta seção são introduzidos conceitos fundamentais para o entendimento da proposta apresentada neste artigo. A seção 2.1 descreve a Ciência dos Dados. A seção 2.2 trata do *Big Data*. A Descoberta de Conhecimento em base de dados é apresentada na seção 2.3. Ainda nesta seção é abordada a Mineração de Dados, Técnica de Classificação e a Técnica de Clusterização. A seção 2.4 apresenta os Trabalhos correlatos.

2.1 CIÊNCIA DOS DADOS

A ciência de dados é uma interseção da ciência da computação e da estatística, integrando a estrutura de dados, algoritmos, sistemas e linguagens de script, bem como um conhecimento sólido de correlação, causalidade e conceitos relacionados que são essenciais para modelar exercícios envolvendo dados. Ela está procurando descobrir conhecimento discutível a partir de uma quantidade grande de dados, que pode ser usada para tomar decisões e fazer previsões, e não simplesmente a interpretação de números. (CONWAY, 2010). A Figura 1 apresenta o ciclo de vida do processo da ciência de dados.

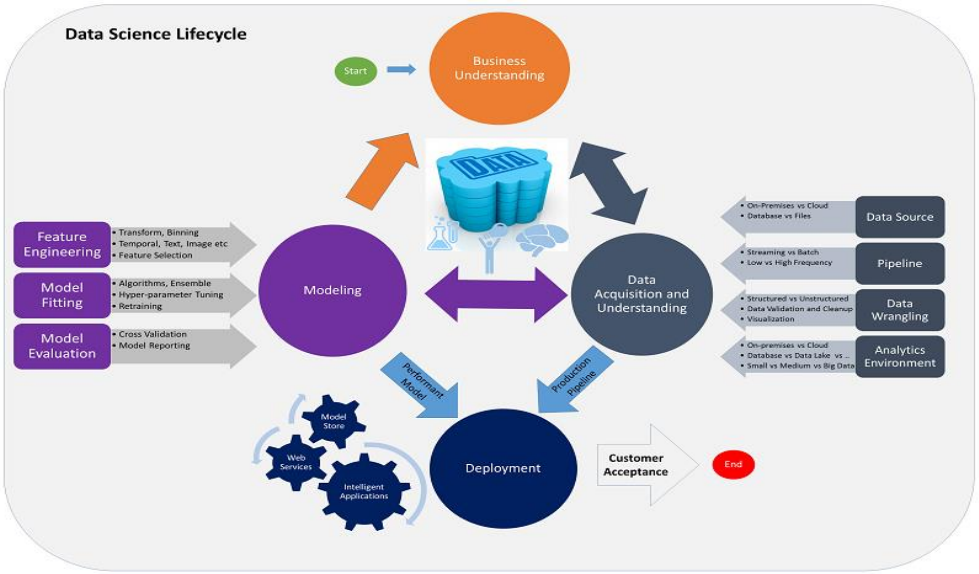

Figura 1 – Ciclo de vida do processo da ciência de dados

Fonte: (CONWAY, 2010)

Conforme ilustra a Figura 1, o ciclo de vida integra as seguintes etapas: noções básicas sobre o domínio da aplicação, aquisição de dados e reconhecimento, modelagem, implantação e aceitação do cliente. Percebe-se que o ciclo de vida do processo é modelado como uma sequência de etapas iteradas que fornecem diretrizes sobre as tarefas necessárias para usar os modelos preditivos. Esses modelos podem ser implantados em um ambiente de produção para serem utilizados a fim de criar aplicativos inteligentes. Neste sentido, a ciência de dados envolve

princípios, processos e técnicas para compreender fenômenos através da análise automatizada de dados (PROVOST; FAWCETT, 2013). Pode-se complementar a definição de ciência de dados com o conceito definido pelos autores (PAIXÃO, SILVA, TANAKA, 2015): "é o domínio científico dedicado para descobrir conhecimento (knowledge discovery) através da análise de dados". Conforme esses autores, os termos atuais, utilizados para definir a Ccência de dados integram outras áreas de conhecimento, tratando-se de um campo interdisciplinar, como: Análise de Dados, Processamento de Dados, Estatística, Descoberta de Conhecimento em Banco de Dados (KDD), Mineração de Dados (Data Mining), Big Data, entre outros. (PAIXÃO, SILVA, TANAKA, 2015), (MATTMANN, 2013) e (VASANT, 2013).

2.2 BIG DATA

Existem diversas definições na literatura para o termo big data. Tomando como base essas definições, big data refere-se à percepção e compreensão de relações entre dados que, até recentemente havia dificuldade para entendimento.

Lima Junior (2012, p. 211) define big data como sendo:

> um [...] conjunto de dados (dataset) cujo tamanho está além da habilidade de ferramentas típicas de banco de dados em capturar, gerenciar e analisar. A definição é intencionalmente subjetiva e incorpora uma definição que se move de como um grande conjunto de dados necessita ser para ser considerado um *big data*.

O autor Davenport (2014) descreve big data como um termo genérico para dados que não podem ser contidos nos repositórios usuais, ou seja, refere-se a dados volumosos demais para caber em um único servidor, não estruturados para se adequar a um banco de dados organizado em linhas e colunas.

Os autores ManyKa et.al (2011) descrevem big data como conjuntos de dados cujo tamanho é além da capacidade de ferramentas de software de banco de dados típicos para capturar, armazenar, gerenciar e analisar.

Já para o autor Canary (2013) big data é definido como ativos de alto volume, velocidade e variedade de informações que exigem custo - benefício, de formas inovadoras de processamento de informações para maior visibilidade e tomada de decisão.

Conforme o autor Mayer (2013), o big data relaciona-se com três importantes mudanças: capacidade de analisar grandes quantidades de dados, aceitar a real confusão dos dados, ao invés de privilegiar a exatidão e respeito por correlação do que pela contínua busca pela causalidade elusiva.

Para Shiffrin (2016), "big data consiste na tarefa de encontrar padrões em grandes volumes de dados". Pode-se observar nas definições descritas pelos pesquisadores que todos se referem a questão da grande quantidade de dados.

A Figura 2 mostra a relação da tomada de decisão presente no big data, podendo assim receber influencia das dimensões volume, variedade, velocidade, valor e veracidade de dados.

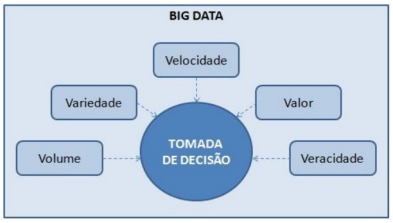

Figura 2 - Tomada de decisão inserida no big data
Fonte: (CANARY, 2013)

Um importante componente que possibilita o gerenciamento e análise do big data é uma nova tecnologia. O autor Davenport (2014) cita as seguintes tecnologias de big data, descritas em maior detalhe na seção 3: Hadoop, MapReduce, Linguagens de Script, Aprendizado de máquina, Visual Analytics, Processamento de linguagem natural e In – Memory Analiytcs.

2.3 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

A pesquisa desenvolvida pelos autores (GDS; PUBLISHING, 2008) comprova que 85 % (oitenta e cinco por cento) de toda a informação do mundo está em formato textual. Analisar base de dados não estruturadas como textos torna-se bastante difícil de sistematizar devido a necessidade de gerar significado do que está escrito. Neste contexto, torna-se necessário a utilização de mecanismos que visem descobrir padrões e informações até então desconhecidos. O KDD (Knowledge Discovery in Database) é um processo complexo que abrange várias etapas relacionadas ao processo de descoberta de conhecimento em base de dados, entre elas o planejamento das atividades (BOENTE; GOLDSCHMIDT; ESTRELA, 2008). De acordo com o autor Goldschimidt e Passos (2005), KDD é um processo, não trivial, interativo, e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. A Figura 3 ilustra as etapas deste processo.

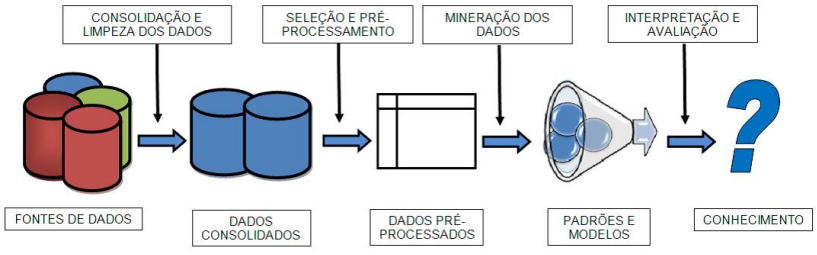

Figura 3 – Etapas do KDD
Fonte: (FIGUEIRA, 1998)

Conforme ilustra a Figura 3, a etapa inicial envolve a consolidação e limpeza dos dados. Com os dados já consolidados, passa-se a etapa da seleção e pre – processamento. Após, aplicação da

técnica de mineração de dados com a geração de padrões e modelos. A etapa final integra a interpretação e avaliação para a geração de conhecimento.

2.3.1 MINERAÇÃO DE DADOS

Mineração de dados (Data Mining) consiste em uma etapa no processo de descoberta de conhecimento em banco de dados (Knowledge Discovery in Database - KDD). A mineração de dados objetiva a análise de grandes conjuntos de dados a fim de encontrar relacionamentos, padrões, tendências de resumir esses dados de uma forma que sejam úteis e possam auxiliar as tomadas de decisões dos mais diversos setores, (HAND et al. 2001).

O processo de análise envolve técnicas matemáticas, estatísticas e computacionais. Processos ou tarefas no domínio de mineração de dados visam construir modelos eficientes capazes de analisar e extrair conhecimento além de predizer tendências futuras no comportamento dos dados.

O autor Weiss (2005, p.87) define Mineração de Dados como:

> Busca de informação valiosa em grandes volumes de dados. É o esforço desenvolvido por homens e máquinas. Os homens desenham os bancos de dados, descrevem os problemas e setam os objetivos. As máquinas mineram os dados, em busca de padrões que atendam a estes objetivos.

Conforme o autor Han (2011), dentre as tarefas mais comuns de mineração de dados é possível citar: Classificação, Agrupamento, Associação e Regressão.

2.3.1.1 TÉCNICA DE CLASSIFICAÇÃO

A técnica de classificação utiliza aprendizado supervisionado. Nesta tarefa os atributos do conjunto de dados são divididos em dois grupos, ou seja, um dos grupos apresenta somente um atributo (atributo alvo) para o qual se deve fazer a predição de um valor. Quando se aplica a técnica de classificar, o atributo alvo é categórico. O outro grupo apresenta os atributos a serem utilizados na predição (GOLDSCHMIDT, PASSOS, BEZERRA, 2015).

O conhecimento descoberto é frequentemente representado na forma de regras SE -> ENTÃO. Essas regras são representadas da seguinte maneira: “SE os atributos preditivos de uma tupla satisfazem as condições no antecedente da regra, ENTÃO a tupla tem a classe indicada no consequente da regra” (MARKOV; LAROSE, 2007).

Os algoritmos de classificação consistem, basicamente, em produzir um modelo de classificação, denominado classificador, a partir de um conjunto de registros existentes, para que, posteriormente, esse modelo seja utilizado para classificar outros exemplos de classe desconhecida (CARVALHO, 2008). A seguir é apresentado o algoritmo J48.

O algoritmo J48 surgiu da necessidade de recodificar o algoritmo C4.5, que, originalmente, é escrito na linguagem C, para a linguagem Java (WITTEN et al., 2005). Ele tem a finalidade de gerar uma árvore de decisão baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste. Um dos aspectos para a grande utilização do algoritmo J48 pelos especialistas em Data Mining é que o mesmo se mostra adequado para os procedimentos, envolvendo as variáveis (dados) qualitativas contínuas e discretas

presentes nas bases de dados.

O algoritmo J48 foi proposto por Quinlan (1993). Para a montagem da árvore, o algoritmo J48 utiliza a abordagem de dividir-para-conquistar, onde um problema complexo é decomposto em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe (MOZZAQUATRO, LIBRELOTTO, 2013). O pseudocódigo para construir a árvore de decisão é descrito no Quadro 1.

Quadro 1 – Pseudocódigo da árvore de decisão

1. Verificar por casos básicos
2. Para cada atributo *a*
    1. Encontrar a razão do ganho da informação normalizado pelo particionamento em *a*
3. Seja *a_maior* o atributo com maior ganho da informação normalizado
4. Criar um *nó* de decisão que particiona o conjunto de dados em *a_maior*
5. Repetir nos subconjuntos obtidos através da divisão em *a_maior*, e adicionar aqueles nós como filhos de *nó*

Fonte: (CARVALHO, 2008)

A Figura 4 ilustra um exemplo de árvore de decisão, apresentando como solução “ir jogar ou não tênis”.

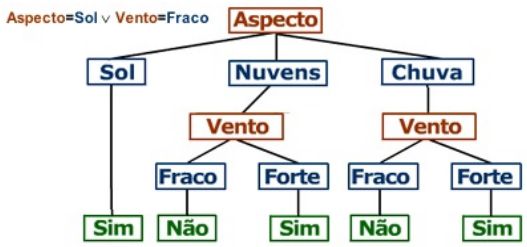

Figura 4– Exemplo de árvore de decisão

Fonte: Adaptado de (GOLDSCHMIDT, PASSOS, BEZERRA, 2015).

2.3.1.2 TÉCNICA DE CLUSTERIZAÇÃO

A técnica de clusterização também é chamada de agrupamento ou análise de agrupamento, utiliza aprendizado não supervisionado. Objetiva-se separar os registros de um conjunto de dados em subconjuntos ou grupos (clusters) de tal forma que elementos em um cluster compartilhem um conjunto de propriedades comuns que os distinguem dos elementos de outros clusters) (GOLDSCHMIDT, PASSOS, BEZERRA, 2015).

Neste sentido a tarefa de agrupar pode ser interpretada como um problema de otimização, ou seja, objetiva-se maximizar a similaridade intracluster e minimizar a similaridade intercluster. Existem na literatura diversos algoritmos de clusterização, neste artigo será abordado o algoritmo

k-means.

O algoritmo de clusterizacao, K-means (também chamado de K-Médias) fornece uma classificação de informações de acordo com os próprios dados. O algoritmo K-means é popular devido a sua facilidade de implementação e sua ordem de complexidade O(n), onde n é o número de padrões. (FONTANA; NALDI, 2009).

K-means utiliza o conceito de centróides como protótipos representativos dos grupos, onde o centróide representa o centro de um grupo, sendo calculado pela média de todos os objetos do grupo. O processo iterativo termina quando os centroides dos grupos param de se modificar, ou após um número preestabelecido de iterações ter sido realizadas.

O algoritmo de K-means pode ser visto através dos passos apresentados no Quadro 2 (FONTANA; NALDI, 2009):

Quadro2 – Passos para execução do algoritmo *K-means*

1. Atribuem-se valores iniciais para os protótipos seguindo algum critério, por exemplo, sorteio aleatório desses valores dentro dos limites de domínio de cada atributo; 2. Atribui-se cada objeto ao grupo cujo protótipo possua maior similaridade com o objeto; 3. Recalcula-se o valor do centróide (protótipo) de cada grupo, como sendo a média dos objetos atuais do grupo; Repete-se os passos 2 e 3 até que os grupos se estabilizem.

Fonte: (FONTANA; NALDI, 2009)

A Figura 5 ilustra o diagrama de atividades do algoritmo *K-means*.

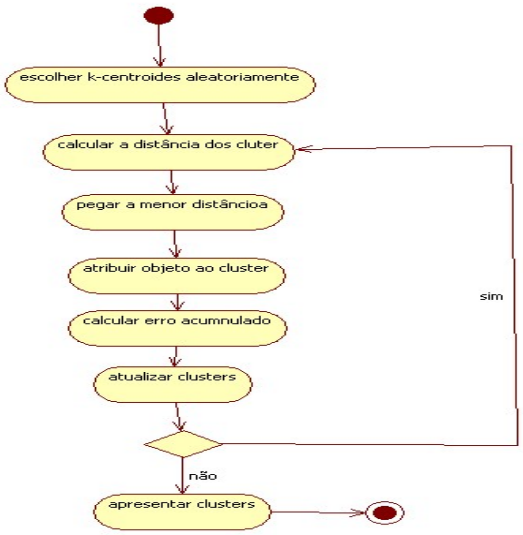

Figura 5 – Diagrama de atividades do algoritmo *K-means*

Fonte: Adaptado de (GOLDSCHMIDT, PASSOS, BEZERRA, 2015)

A técnica de classificação diferencia-se da clusterização, pois na classificação os registro estão associados a rótulos predefinidos. Já no agrupamento, os objetos considerados como entrada não possuem rótulos associados. A clusterização analisa grupos por meio de um conjunto de objetos em grupos de acordo com alguma medida de similaridade.

2.4 TRABALHOS CORRELATOS

Esta seção é dedicada à apresentação dos resultados de pesquisas realizadas por autores sobre ciência de dados, ciclo de vida dos dados, big data e mineração de dados. A análise desses trabalhos serviu de fundamentação teórica para embasamento desta pesquisa.

A pesquisa desenvolvida por Júnior (2003) intitulada "Desenvolvimento de um Framework para análise visual de informações suportando Data Mining" objetivou a construção de um arcabouço de visualização de dados que intenciona potencializar o processo KDD. O trabalho combinou uma série de técnicas de visualização integradas. O resultado refere-se a soma das principais vantagens de cada uma das abordagens. É construída uma ferramenta para o aproveitamento de dados e geração de conhecimento.

A pesquisa desenvolvida por Metz (2006) intitulada "Interpretação de clusters gerados por algoritmos de clustering hierárquico" propõe o desenvolvimento de um módulo de aprendizado não supervisionado, que agrega algoritmos de clustering hierárquico e ferramentas de análise de clusters para auxiliar o especialista na interpretação dos resultados. Para avaliar o módulo proposto e seu uso na descoberta de conhecimento a partir da estrutura de clusters foram realizados diversos experimentos. Os resultados mostram a viabilidade da metodologia proposta para interpretação dos clusters, apesar da complexidade do processo ser dependente das características do conjunto de dados.

O estudo realizado por Bufrem et.al (2016) intitulado de “Produção Internacional Sobre Ciência Orientada a Dados: análise dos termos Data Science e E-Science na Scopus e na Web of Science” utiliza métodos bibliométricos para analisar a produção científica sobre o tema. Realiza uma busca nas bases de dados Web of Science (WoS) e Scopus, limitada ao período mais profícuo para o tema que corresponde ao horizonte temporal entre 2006 e 2016 para identificar os estudos, autores, periódicos e temáticas mais proeminentes, categorizando-os e relacionando-os, com vistas à identificação de sua constelação temática. Como resultado, foi possível observar que até 2013 havia uma grande preocupação dos pesquisadores com as questões ligadas à computação em grade (Grid computing), mostrou-se os temas em destaque nos tópicos “e-Science” e “Data Science” na Scopus e na WoS. A partir de 2014 o foco dos trabalhos esteve mais ligado ao big data. Foram também analisadas as relações dos temas entre si, entre autores e temas, periódicos e temas, tanto na base Scopus quanto na base WoS e analisados os temas mais representativos, das duas bases estudadas, considerando apenas as áreas de Biblioteconomia e Ciência da Informação (BCI). Sobre os periódicos mais representativos, destacou-se a área de Tecnologia da Informação (TI), com os periódicos Future Gener Comput Syst (destaque na Scopus) e Concurr Comp- Pract (destaque na WoS). Com a pesquisa percebeu-se que a área e-Science e Data Science tem-se destacado.

Sant’Ana (2016) em seu estudo intitulado “Ciclo de vida dos dados: uma perspectiva a partir

da Ciência da Informação" abordou uma proposta de um novo olhar para o Ciclo de Vida dos Dados, que pressupõe, como elemento central, os próprios dados, amparando-se nos conceitos e contribuições que a Ciência da Informação pode proporcionar, sem abrir mão da reflexão sobre o papel de outras áreas chave como a Ciência da Computação. Como resultados apresentam-se as fases de coleta, armazenamento, recuperação e descarte, permeadas por fatores transversais e presentes em todas as fases: privacidade, integração, qualidade, direito autoral, disseminação e preservação, compondo um Ciclo de Vida dos Dados. Constatou que big data requer novos olhares para os processos de acesso e uso de dados. A Ciência da Informação pode oferecer um novo enfoque, agora centrado nos dados, e contribuir para a otimização do Ciclo de Vida dos Dados como um todo, ampliando as pontes entre os usuários e os dados que necessitam.

"Curadoria Digital: proposta de um modelo para Curadoria Digital em ambientes big data baseado numa abordagem semi-automática para a seleção de objetos digitais", desenvolvida por Dutra et.al (2016), propõe técnicas de seleção e avaliação de objetos digitais para curadorias digitais que levem em conta o volume, a velocidade, a variedade, a veracidade e o valor dos dados coletados em múltiplos domínios do conhecimento. Apresenta um modelo para busca, tratamento, avaliação e seleção de objetos digitais a serem tratados em curadorias digitais.

A pesquisa intitulada "A formação em ciência de dados: uma análise preliminar do panorama estadunidense", de Curty e Serafim (2016), procura caracterizar e compreender os aspectos formativos do cientista de dados. O artigo relata um recorte de uma pesquisa de levantamento com base na análise preliminar de 93 cursos em ciência de dados ofertados por instituições estadunidenses. A análise de conteúdo das informações contidas nos websites dos programas identificados permitiu evidenciar que este profissional é formado para lidar com aspectos relacionados à coleta, tratamento, transformação, análise, visualização e curadoria de grandes e heterogêneas coleções de dados orientadas à resolução de problemas práticos e reais. Foi possível constatar que, de modo geral, a formação em ciência de dados atribui grande ênfase a habilidades estatísticas, matemáticas e computacionais, incluindo programação e modelagem avançada.

3. METODOLOGIA

Nesta seção serão descritos os procedimentos metodológicos adotados para o desenvolvimento do trabalho. Primeiramente são apresentadas as características da pesquisa, a fim de identificar e classificar este estudo. Em seguida são apresentadas as etapas, com o objetivo demonstrar e explorar seu contexto. A seção também aborda os instrumentos (softwares e hardwares) necessários para desenvolver a proposta, bem como os instrumentos necessários para realizar os testes.

Do ponto de vista de seus objetivos, conforme descreve Gil (2002), Marconi e Lakatos (2003), este estudo é classificado como pesquisa exploratória, pois busca por maior familiaridade acerca de um problema de pesquisa, tornando-o explícito. Em geral, este trabalho assume as formas de pesquisas bibliográficas, questionamentos realizados e estudos de casos (GIL, 2008; MARCONI e LAKATOS, 2003).

A partir da definição das características da pesquisa, se delimitou as etapas necessárias para o desenvolvimento do estudo. Este estudo será executado nas seguintes fases: Fase I – Insights sobre o Trabalho e Aspectos Teóricos; Fase II – Modelagem da proposta; Fase III – Testes propostos; Fase IV – Conclusões.

Na primeira fase foi desenvolvido o estudo teórico, o qual buscou fundamentos, estratégias e características importantes sobre a Ciência de Dados, Big Data, Descoberta de conhecimento em base de dados, Mineração de dados, Técnicas de classificação e clusterização, como também os algoritmos integrantes das mesmas: J48 e K –means. A partir da revisão bibliográfica, foram definidos os objetivos concretos para a construção da proposta de um modelo híbrido baseado em classificação e clusterização para a extração de informação a partir de grandes bases de dados. A segunda fase consiste na descrição da proposta. O fluxo de um processo foi desenvolvido na linguagem UML (PRESSMAN, 2011). Foi construído um diagrama de sequencia com a ferramenta StarUML . A terceira fase constitui na definição de recursos de infraestrutura tecnológica, bem como a instalação dos softwares necessários para o desenvolvimento da proposta.

Como recursos de software cita-se a linguagem PHP, utilizando a aplicação DreamWeaver para o desenvolvimento e a base de dados MySQL para o armazenamento das informações colhidas no ambiente, utilizando a plataforma Xamp, também será necessário a utilização das seguintes tecnologias: Hadoop, MapReduce, Linguagens de Script, Aprendizado de máquina, Visual Analytics, Processamento de linguagem natural e In – Memory Analiytcs, software Waikato Environment for Knowledge Analysis (WEKA) e SQL Server Integration Services (SSIS).

A quarta fase integra a proposta de aplicação dos testes. Os testes iniciais poderão ser realizados com o software minerador Waikato Environment for Knowledge Analysis (WEKA).

3.1 PROPOSTA DE UM MODELO HÍBRIDO BASEADO EM CLASSIFICAÇÃO E CLUSTERIZAÇÃO PARA A EXTRAÇÃO DE INFORMAÇÃO A PARTIR DE GRANDES BASES DE DADOS

A Figura 6 ilustra um exemplo, descrito por meio de um diagrama de sequencia, o processo fluxo de pedidos em um contexto empresarial.

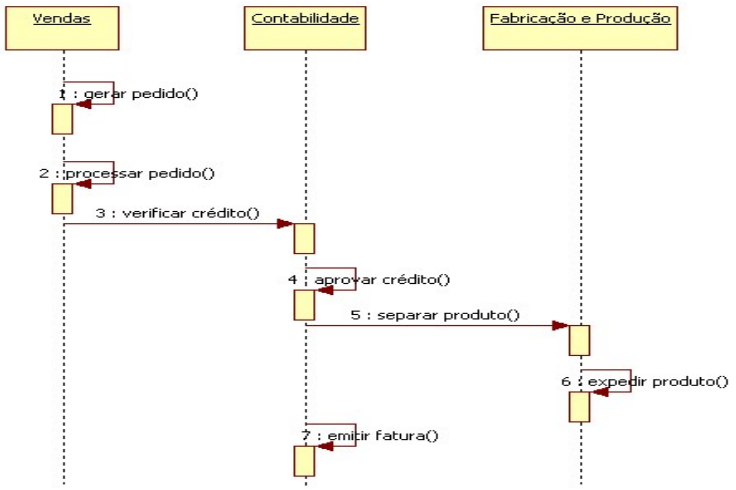

Figura 6 – Processo Fluxo de Pedidos

Fonte: Elaborado pelo Autor

Conforme observa-se na Figura 6, existe um módulo Vendas, Contabiliadde e Fabrição. As vendas geram os pedidos, para após serem apersentados. A contabilidade verifica o crédito, poderá ou não aprová-lo. Caso aprovado emite fatura. O setor responsável pela fabricação separa os produto e após expede-os. O processo citado mostra a interação entre os módulos Vendas, Contabilidade e Fabricação. Quando se trabalha com big data (grandes bases de dados), existem dados em diversificados formatos, por exemplo documentos e pdfs, logs, imagens, vídeos, mídias sociais dentre outros.

Pode-se observar que as fontes de dados são heterogêneas envolvendo um conjunto de dados não estruturados e semiestruturados mais diversificados. As empresas que adotam ambientes de big data necessitam de maneira mais rápida processar grandes volumes de dados. A tecnologia Hadoop pode ser utilizada a fim de ingerir rapidamente os dados, processá-los e armazená-los para utilização. Esta tecnologia possui algoritmos analíticos avançados para orientar previsões e tomadas de decisão.

A proposta aqui apresentada indica a utilização das seguintes tecnologias, conforme o autor Davenport (2014):

1- Hadoop: é um software de código aberto para o processamento de big data em uma série de servidores paralelos;
2- MapReduce: é um framework no qual o Hadoop se baseia;
3- Linguagens de Script: linguagens de programação adequada ao big data;
Aprendizado de máquina: software para identificar o modelo mais adequado ao conjunto de dados;
4- Visual Analytics: os resultados analíticos são apresentados em formato visual ou gráfico;
5- Processamento de linguagem natural: software para análise de texto;
6- In – Memory Analiytcs: processamento de big data na memória do computador a fim de obter maior velocidade.

Para a realização da integração entre os diversificados formatos de dados será utilizado o SQL Server Integration Services (SSIS), o mesmo possui uma combinação de tecnologias heterogêneas para a manipulação de dados (tais como bancos de dados relacionais, XML, arquivos texto, planilhas do Excel, dentre outros) a recursos do .NET Framework, esta ferramenta possibilita a construção de aplicações capazes de se comportar de uma forma robusta e escalável grandes volumes de dados. (PAULA, 2016). Por meio de sua implementação não é necessário substituir os sistemas existentes e, sim fazer com que eles troquem dados de forma transparente ao usuário, causando a impressão que é um único sistema.

Os serviços serão acessados em forma de aplicações, que podem ser acessados através de uma URL. Esses serviços fazem os sistemas conversarem uns com os outros, independente da linguagem de programação que foram construídos.

Para que se possa gerar conhecimento em relação a fontes de dados heterogêneas com dados não estruturados e semiestruturados diversificados, torna-se necessário criar uma nova base homogênea com dados em formato único tornando-se assim legíveis ao sistema.

Com a nova base gerada, serão aplicadas técnicas de Mineração de dados a fim de gerar conhecimento. As seguintes etapas serão implementadas:

Etapa 1 - consolidação e limpeza dos dados;
Etapa 2 - seleção e pré-processamento;

Etapa 3 - aplicação da Mineração de Dados. Será criado um framework híbrido integrando as técnicas de classificação e clusterização. A classificação irá gerar uma árvores de decisão por meio da aplicação do algoritmo J48, utilizando o aprendizado supervisionado. Também será aplicado o algoritmo k-means, a fim de realizar um agrupamento dos dados similares, integrando o aprendizado não supervisionado.

Etapa 4 – Geração de padrões e modelos a fim de inferir conhecimento. A tomada de decisão poderá avaliar as seguintes variáveis: veracidade, volume, velocidade e valor da informação gerada.

4-CONSIDERAÇÕES

As organizações empresariais buscam alternativas para melhorar seus processos de negócio. Uma delas é integrar suas aplicações a fim de que funcione de forma sincronizada e eficiente. Neste contexto, torna-se necessário a utilização de mecanismos computacionais que permitam a interação com estes dados de forma inteligente e rápida, assim surgindo a ciência de dados. Integrar aplicações é fazer com que diferentes aplicações que não foram concebidas tendo em mente sua integração, possam colaborar para dar suporte a um novo processo de negócio.

Assim, a pesquisa aqui apresentada tem por objetivo gerar uma base de dados única integrando um conjunto de aplicações heterogêneas e aplicar técnicas de mineração de dados na mesma a fim de gerar conhecimento. A pesquisa aborda dois processos:

A etapa um trata da geração de uma base de dados única partindo de um conjunto de aplicações heterogêneas em diversificados formatos. Com o estudo realizado propõe-se utilizar o SQL Server Integration Services (SSIS), pois possui uma combinação de tecnologias heterogêneas para a manipulação de dados (bancos de dados relacionais, XML, arquivos texto, planilhas do Excel) e as seguintes tecnologias: Hadoop, MapReduce, Linguagens de Script, Aprendizado de máquina, Visual Analytics, Processamento de linguagem natural e In – Memory Analiytcs.

A etapa dois envolve a aplicação de técnicas de Mineração de dados para inferir informações desses dados, ou seja, como gerar conhecimento a partir de uma grande base de dados para uma tomada de decisão. Os testes iniciais serão realizados com o software Waikato Environment for Knowledge Analysis (WEKA) implementando os algoritmos J48 (técnica de classificação) e K-means (técnica de clusterização).

REFERÊNCIAS

AGARWAL, R., VASANT, D. **Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research**. Information Systems Research 25(3):443448, 2014

BOENTE, Alfredo Nazareno Pereira; GOLDSCHMIDT, Estrela; VIEIRA, Vania.. **Uma Metodologia de Suporte ao Processo de Descoberta de Conhecimento em Bases de Dados**. In: V Simpósio de Excelência em Gestão e Tecnologia, 2008, Resende - RJ. V SEGeT, 2008.

BUFREM, Leilah Santiago; SILVA, Fábio Mascarenhas; SOBRAL, Natanael Vitor; CORREIA, Anna Elizabeth Galvão Coutinho. **Produção internacional sobre ciência orientada a dados: análise dos termos Data science e e-science na scopus e na Web of science**. Revista Informação e Informação. DOI: 10.5433/1981-8920.2016v21n2p40. Londrina, v. 21, n. 2, 2016, p. 40 – 67.

Disponível em: < http:www.uel.br/revistas/informacao/>. Acesso em mai de 2017.

CANARY, Vivian Passos. **A tomada de decisão no contexto do big data: Estudo de caso único**. Universidade Federal do Rio Grande do Sul. Escola de Administração Trabalho de Conclusão de Curso. Porto Alegre, 2013

CARVALHO, D. D.; DIAS, M. M. **Descoberta de Conhecimento em Ambientes Virtuais de Aprendizagem: Um Estudo de Caso no LabSQL**. Trabalho de Conclusão de Curso em Ciência da Computação. Universidade Federal do Pará, Belém, 2008

CONWAY, D. **The Data Science Venn Diagram**, 2010. Disponível em Acesso em mai de 2017

CURTY, Renata Gonçalves; CERVANTES, Brígida Maria Nogueira. **Data Science: ciência orientada a dado**s. Departamento de Ciência da Informação – Universidade Estadual de Londrina (UEL). DOI: 10.5433/1981-8920.2016v21n2p40. Londrina, v. 21, n. 2, 2016, p. 1 – 3. Disponível em: < http:www.uel.br/revistas/informacao/>. Acesso em mai de 2017.

CURTY, Renata Gonçalves; SERAFIM, Jucenir da Silva. **A formação em ciência de dados: uma análise preliminar do panorama estadunidense**. Revista Informação e Informação. DOI: 10.5433/1981-8920.2016v21n2p307. Londrina, v. 21, n. 2, p. 307–328. Disponível em: < http:www.uel.br/revistas/informacao/>. Acesso em mai de 2017.

DAVENPORT, T. H., **Big Data at Work: Dispelling the Myths**, Uncovering the Opportunities, Harvard Business School Publishing Corporation, 2014.

DAVENPORT, Thomas. H., **Big Data no trabalho: Derrubando mitos e descobrindo oportunidades.** Edição 1, Rio de Janeiro, editora Elsevier, 2014.

DAVENPORT, Thomas. H., **Dados demais: como desenvolver habilidades analíticas para resolver problemas complexos, reduzir riscos e decidir melhor**. Edição 1, Rio de Janeiro, editora Elsevier, 2014, 240 p.

DUTRA, Moisés Lima; MACEDO, Douglas Dyllon Jeronimo de. **Curadoria digital: proposta de um Modelo para curadoria digital em Ambientes big data baseado numa Abordagem semi-automática para a Seleção de objetos digitais**. Revista Informação e Informação. DOI: 10.5433/1981-8920.2016v21n2p143. Inf. Inf., Londrina, v. 21, n. 2, p. 143 – 169. Disponível em: < http:www.uel.br/revistas/informacao/>. Acesso em mai de 2017.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

FONTANA, A., Naldi, M. C. **Estudo de Comparação de Métodos para Estimação de úmeros de Grupos em Problemas de Agrupamento de Dados**. Universidade e São Paulo. ISSN - 0103-2569. 2009. Disponivel em: Acesso em mai de 2017.

GANTZ, J. e REINSEL, D. **Extracting Value from Chaos,** IDC iView, 2011 Disponível em

http://www.emc.com/digital_universe..www.emc.com/collateral/analyst-reports/idcextracting-value-from-chaos-ar.pdf. Acesso em mai de 2017

GDS PUBLISHING. **Managing the Data Explosion**. Business Management,2008.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. - São Paulo : Atlas, 2002

GOLDSCHIMIDT, R ; PASSOS, E. **Data mining: Um guia prático**. Rio de Janeiro: Campus, 2005.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações. 2.ed,. Rio de Janeiro: Elsevier, 2015.

GRUS, Joel. **Data Science do Zero.** Rio de Janeiro: Alta Books, 2016. CAPITULO 1

HAN, J. M. Kamber; J. Pei. **Data mining: concepts and techniques: concepts and techniques**. Elsevier, 2011.

JÚNIOR, José Fernando Rodrigues. **Desenvolvimento de um framework para análise visual de informações suportando Data Mining**. Dissertação de Mestrado. Instituto de Ciências Matemática e de Computação. USP. São Carlos, 2003.

LIMA Junior, W. T. **Big Data, Jornalismo Computacional e Data Journalism: estrutura, pensamento e prática profissional na Web de Dados**. Estudos em Comunicacao, v. 12, p. 207-222, 2012

MARKOV, Z.; LAROSE, D. **Data Mining The Web: Uncovering Patterns in Web Content, Structure, and Usage**. Published by John Wiley & Sons, Inc., Hoboken, NewJersey, 2007.

MATTMANN, C. A. **Computing: A vision for data science**, Nature, 493, 473–475, 2013.

METZ, Jean. **Interpretação de clusters gerados por algoritmos de clustering hierárquicos.** Dissertação de mestrado em ciência da computação e matemática computacional. São Paulo: Instituto de Ciências Matemáticas e de Computação – USP, 2006

HAND, D; MANNILA, H; SMYTH, P. **Principles of Data Mining**. MIT Press, 2001. Disponível em: < http://www.ru.lv/~peter/zinatne/ebooks/DataMining1.pdf>. Acesso em mai de 2017

LAKATOS, Eva Maria. MARCONI, Marina de Andrade. **Fundamentos de Metodologia Científica.** 5 ed. São Paulo: Atlas 2003.

MANIYK A, James; et. al. **Big data: The next frontier for innovation, competition, and productivity**. McKinseyGlobalInstitute, 2011. Disponível em:http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation. Acesso em mai de 2017.

MAYER, Schonberger Viktor. **Big Data: como extrair volume, variedade, velocidade e valor**

da avalanche de informação cotidiana. 1.ed, Rio de Janeiro: Elsevier, 2013

MOZZAQUATRO, Patricia Mariotto; LIBRELOTTO, Solange Rubert. **Análise dos algoritmos de mineração j48 e apriori aplicados na detecção de indicadores da qualidade de vida e saúde** .Revint. Revista interdisciplinar de ensino pesquisa e extensão, 2013, v1.

PAIXÃO, Alexandre de Oliveira ; SILVA, Verônica Aguiar da; TANAKA, Asterio. **De Business Intelligence a Data Science: um estudo comparativo entre áreas de conhecimento relacionadas**. Congresso integrado da tecnologia da informação, 2015.

PALETTA, Francisco Carlos. **A informação e a biblioteconomia o perfil profissional na era da web**, XVIII Seminário Nacional de Bibliotecas Universitárias, SNBU 2014. Belo Horizonte

PAULA, Danieli. **Business intelligence no auxílio da gestão da Inovação: um estudo de caso utilizando SQL SERVER** Integration services e microstrategy. Revista Interdisciplinar Científica Aplicada, Blumenau, v.10, n.2, p.69-92, TRII 2016. ISSN 1980-7031.

PRESSMAN, Roger S. **Engenharia de software: uma abordagem profissional**. 7.ed.. PORTO ALEGRE: AMGH, 2011. 780 p. ISBN 978-85-63308-33-7.

PROVOSt, F., Fawcett, T. "**Data Science and its relationship to big data and data-driven decision making**" In: Big Data Journal, Vol. 1, pp. 51-59, 2013.

QUINLAN, J. R.; C4.5: **Programs for machine learning.Morgan Kaufmann Publishers** Inc., San Francisco, CA, USA, 1993

SANT'ANA, Ricardo César Gonçalves. **Ciclo de vida dos dados: uma Perspectiva a partir da ciência da Informação. Revista Informação e Informação**. DOI: 10.5433/1981-8920.2016v21n2p116. Londrina, v. 21, n. 2, p. 116 – 142. Disponível em:<.http:www.uel.br/revistas/informacao/>. Acesso em mai de 2017.

SHIFFRIN, Richard M. **Drawing causal inference from big data**. Proceedings of the National Academy of Sciences, Washington, v. 113, n. 27, p. 7308-7309, 2016

VASANT, D. **Data Science and Prediction**. Communications of the ACM, vol. 56, no. 12, 2013.

WEISS, S. M., Indurkhya, N., Zhang, T., e Damerau, F. **Text Mining: Predictive Methods for Analyzing Unstructured Information**. Springer S cience+Business Media, Inc, p. 1-100, 2005.

WITTEN, I. H.; FRANK, E. **Data Mining: practical machine learning tools and techniques with java implementations**. 2.ed. Morgan Kaufmann, San Francisco, CA, 2005.