UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DIOVAN OLIVEIRA LEAL

ANÁLISES IMPLÍCITAS DE DADOS NA PRODUÇÃO DE CONHECIMENTO EM CIÊNCIA DA COMPUTAÇÃO: UM ESTUDO BIBLIOMÉTRICO

CRICIÚMA
2019

DIOVAN OLIVEIRA LEAL

ANÁLISES IMPLÍCITAS DE DADOS NA PRODUÇÃO DE CONHECIMENTO EM CIÊNCIA DA COMPUTAÇÃO: UM ESTUDO BIBLIOMÉTRICO

Trabalho de Conclusão de Curso, apresentado para obtenção do grau de Bacharel no curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientador: Prof. Dr. Kristian Madeira
Coorientadora: Profª. Dra. Merisandra Côrtes de Mattos Garcia

CRICIÚMA

2019

DIOVAN OLIVEIRA LEAL

ANÁLISES IMPLÍCITAS DE DADOS NA PRODUÇÃO DE CONHECIMENTO EM CIÊNCIA DA COMPUTAÇÃO: UM ESTUDO BIBLIOMÉTRICO

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Inteligência Artificial.

Criciúma, 27 de junho de 2019.

BANCA EXAMINADORA

Prof. Kristian Madeira - Dr. - (UNESC) - Orientador

Profa. Merisandra Côrtes de Mattos Garcia -Dra. - Coorientadora (UNESC)

Prof. Fabrício Giordani - Esp - (UNESC)

Prof. Paulo João Martins - Me - (UNESC)

À minha Família.

AGRADECIMENTOS

Agradeço primeiramente a Jeová Deus, senhor criador. Aos meus Pais Antonio Enilton Lopes Leal e Iracema Oliveira Leal que mesmo distante sempre me apoiaram e me deram forças para continuar. A minha esposa Alline Pinto de Matos que me apoiou neste momento de muito trabalho e dedicação, sendo necessário sacrificar momentos de sua companhia.

Agradeço ao Dr Kristian Madeira meu orientador, que aceitou o desafio de orientar a realização de todo um trabalho de conclusão de curso em apenas um semestre. Obrigado por suas valorosas contribuições.

Também sou grato à Dra Merisandra Côrtes de Mattos Garcia minha coorientadora que me deu suporte necessário desde a escolha do orientador assim como sua própria coorientação.

Aos membros do Laboratório de Pesquisa Aplicada em Computação e Métodos Quantitativos (LACOM) da UNESC que me deram todo apoio e auxílio na realização desta pesquisa.

Aos citados, meu sincero agradecimento espero sempre honrá-los em meus feitos.

“... Não é sobre chegar no topo do mundo e saber que venceu. É sobre escalar e sentir que o caminho te fortaleceu ...”

Ana Vilela

RESUMO

Ciência e negócios são exemplos de áreas afetadas em decorrência do notável volume e variedade de dados atualmente disponíveis. Com isto uma área de estudos fica em evidência, ciência dos dados. O grande desafio é analisar esta quantidade de dados e gerar informação. Necessitando o emprego de técnicas apropriadas, as análises implícitas. Dada a importância destes algoritmos em nosso cotidiano, as produções científicas fundamentadas nesta área também avultam. Então, pela bibliometria, campo de estudo da ciência da informação, que de forma quantitativa e estatística avalia as produções científicas. Este trabalho tem por objetivo, desenvolver uma pesquisa bibliométrica na ciência da computação a partir de trabalhos que empregam técnicas de análises implícitas. Além do mapeamento bibliométrico, também foi realizada a fundamentação teórica sobre ciência dos dados, análises implícitas e bibliometria. São abordadas as seguintes análises implícitas: Apriori, árvores de decisão, classificadores *bayesianos*, *DBSCAN, FP-Growth,* máquinas de vetores de suporte, redes neurais artificiais, *k-means e k-medoid*. Os artigos científicos analisados são oriundos de três bases de dados, *SciElo*, *Scopus* e *Web of Science*. A pesquisa seguiu os seguintes critérios de inclusão de arquivos : artigos aplicados à computação, utilizar alguma das análises implícitas e não ser uma bibliometria. Ao fim da pesquisa bibliométrica com volume de 46 artigos, dos quais foram obtidos resultados e conclusões relevantes ao cenário da pesquisa de análises implícitas em ciência da computação. Por meio do *h-index,* os três principais autores são: Brázdil Thomáš, Artur S. D'Avila Garcez e Mahajan, Meena com os respectivos *h-index,* quinze, treze e doze, é identificado que o pesquisador Ye, Yongkai destaca-se por ser o único autor com mais de um trabalho nesta pesquisa, assim como, estabelece uma relação de coautoria em demais trabalhos. Ainda, o ano de 2018 foi o ano mais produtivo com dezesseis artigos, também destaca-se China e Índia pelas suas produtividades, nove e sete respectivamente. Também, a partir dos artigos destaca-se cinco grupos de pesquisas: Pesquisa e Desenvolvimento, Processamento de Linguagem Natural, Segurança Computacional, Pesquisa e Indexação de Conteúdo e Ausência de Dados em *datasets.* As análises mais utilizadas foram árvores de decisão, *Apriori* e redes neurais artificiais. De acordo com os resultados obtidos, conclui-se que este campo de pesquisa encontra-se em crescimento, possui pelo menos duas subáreas de tendência de pesquisa: Pesquisa e Desenvolvimento Computacional e Processamento de Linguagem Natural, além de uma lacuna de pesquisa, Ausência de Dados em *datasets.* Ainda, entre os autores, confirma-se a existência de uma relação de cooperação qual é identificado pelos trabalhos do autor Ye, Yongkai e também os estudos apontam para análises mais utilizadas, árvores de decisão, *Apriori* e redes neurais artificiais.

Palavras-chave: Bibliometria. Análises implícitas. Ciência dos dados. Mapeamento bibliométrico.

ABSTRACT

Science and business are examples of affected areas as a result of the volume and variety of data currently available. With this, an area of study stands out, data science. The big challenge is to analyze this big amount of data and generate information. Then the use of appropriate techniques is needed, the implicit analyzes. Given the importance of these algorithms in our daily lives, the scientific productions on this area also increase. Then, through bibliometrics, information science's field of study that quantitatively and statistically evaluates the scientific production. This work aims to develop a bibliometric research in computer science evaluating other works that employ the implicit analysis. Besides the bibliometric mapping, the theoretical basis on data science, implicit analysis and bibliometry was also carried out. The following implicit analyzes are addressed: Apriori, decision trees, Bayesian classifiers, DBSCAN, FP-Growth, support vector machines, artificial neural networks, k-means and k-medoid. The scientific papers analyzed come from three databases, SciElo, Scopus and Web of Science. The research followed the following criteria for inclusion of files: articles applied to computing, using some of the implicit analysis and not being a bibliometry. At the end of the bibliometric research there was a volume of 46 articles, from which were obtained results and conclusions relevant to the scenario of the research of implicit analysis in computer science. By means of h-index, the three main authors are: Brázdil Thomáš, Artur S. D'Avila Garcez and Mahajan, Meena with their respective h-index, fifteen, thirteen and twelve, it is identified that the researcher Ye, Yongkai, is the only author with more than one work in this research, as well as establishing a co-authoring relationship in other works. Also, the year 2018 was the most productive year with sixteen articles, China and India also stand out for their productivities, nine and seven respectively. Also, from the articles stand out five research groups: Research and Development, Natural Language Processing, Computational Security, Content Indexing and Search and Data Absence in datasets. The most used analyzes were decision trees, Apriori and artificial neural networks. According to the results obtained, it is concluded that this field of research is growing, has at least two sub-areas of research trends: Research and Development and Natural Language Processing, in addition to a research gap, Absence of Data in datasets. Also, the authors confirm the existence of a cooperative relationship which is identified by the works of the author Ye, Yongkai and also the studies point to more used analyzes, decision trees, Apriori and artificial neural networks.

Keywords: Bibliometry, Implicit analysis, Data Science, Bibliometric Mapping.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AI	*Article Influence Score*
BG	Busca Gulosa
BRKGA	*Biased Random-Key Genetic Algorithm*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CapesNet	*Deep Neural Network Capsules*
CART	*Classification And Regression Trees*
CHAID	*Chi Square Automatic Interaction Detection*
CI	*Con-ditional Independence Tests*
CLARA	*Clustering Large Applications*
CLARANS	*Clustering Large Applications Based Upon Randomized Search*
CN2	*Induction Algorithm*
CNN	*Convolutional Neural Network*
DAE	*Deep Auto-Encoders*
DBN	*Deep Belief Network*
DBSCAN	*Density Based Spatial Clustering of Application With Noise*
DNN	*Deep Neural Networks*
EF	*Eigenfactor Score*
ETL	*Extract Transformation and Load*
FI	*Fator de Impacto*
FP	*Frequent Pattern*
FP-Growth	*Frequent Pattern-Growth*
FP-Tree	*Frequent Pattern-Tree*
GAN	*General Adversarial Network*
GAN	*General Augmented Naïve Bayes*
IA	Inteligência artificial
ICS	*Inductive Causation Search*

ID3	*Inductive Decision Tree*
ISI	*Institute* for *Scientific Information*
IEEE	*Institute of Electric and Electronic Engineers*
LACOM	Laboratório de Pesquisa Aplicada em Computação e Métodos Quantitativos
LDA	*Latent Dirichlet Allocation*
LSTM	*Long Short-Term Memory*
MPNN	*Multilayers Perceptrons Neural Networks*
NB	*Naïve Bayes*
NER	*Named Entity Recognition*
PLN	Processamento de Linguagem Natural
NOSQL	*Not Only SQL*
PAM	*Partitioning Around Medoids*
PMI	*Point Wise Mutual Information*
RF	*Random Forest*
RNA	Redes Neurais Artificiais
RNN	*Recurrent Neural Network*
SCALPARC	*Scalable Parallel Classifier*
SLIQ	*Supervised Learning In Quest*
SMO	*Sequential Minimal Optimization*
SQL	*Structured Query Language*
SPELL	*Scientific Periodicals Electronic Library*
SPRINT	*Scalable PaRallelizable INduction of Decision Tree*
SVM	*Support Vector Machine*
TAN	*Tree Augmented Naïve Bayes*
TREPAN	*Trees Parroting Networks*
UFSC	Universidade Federal de Santa Catarina

SUMÁRIO

1 INTRODUÇÃO

Atualmente conceitos a respeito dos paradigmas tecnológicos e científicos, teorizados por autores como Halévy (2010) e Schwab (2016), demonstram uma relevância nunca antes observada para os dados como conteúdos essenciais na era do conhecimento e quarta revolução industrial.

A expressiva geração de dados em ampla escala impõe mudanças na forma como as coisas operam, ciência e negócios são exemplos de áreas afetadas em decorrência do notável volume, velocidade e variedade de dados atualmente disponíveis (CURTY; CERVANTES, 2016).

Com isto uma área de estudos fica em evidência, a área da ciência que lida com as tratativas condizentes aos dados é a ciência dos dados (*data science*) ou também ciência orientada a dados, é um campo relativamente novo, abrangente e interdisciplinar, que a vista da explosão dos dados, fenômeno também conhecido como *big data* tornou-se necessariamente conhecida (BUFREM et. al., 2016).

A exposição dos termos referente a ciência dos dados é notória. Publicações sobre ciência dos dados são normais e frequentes, tanto voltadas para os negócios e outras áreas, quanto para ciência (AGARWAL; DHAR, 2014, tradução nossa).

Acompanhando a popularização da ciência dos dados, as produções científicas fundamentadas nesta área também avultam. A necessidade frente ao desafio de extrair dados e gerar informações atrai pesquisadores de muitas áreas do conhecimento (FAGUNDES; MACEDO; FREUND, 2017).

Embora as produções científicas na ciência orientada a dados, possam ser classificadas em variados temas e objetivos, observa-se a predominância de um objetivo macro comum. A capacidade de extrair conhecimento de grandes e heterogêneas bases de dados, por meio de métodos de análises implícitas (BUFREM et. al., 2016).

Os métodos de análises de dados são classificados em explícitos e implícitos, no primeiro caso, a informação é obtida com métodos de baixa complexidade, no segundo caso, ou seja, os métodos implícitos o qual é parte deste

trabalho, a descoberta de conhecimento demanda o uso de métodos mais apurados, como: árvores de decisão, classificadores *bayesianos*, redes neurais artificiais, máquina de vetores de suporte, *k-means*, *k-medoid*, *DBSCAN*, Apriori e *FP-Growth* (AMARAL, 2016).

O entendimento e a contextualização destes métodos se fazem necessários, visto que dependendo do problema a ser resolvido é utilizado um ou outro ou até mesmo uma combinação de alguns destes métodos.

Estes métodos possuem resultados variados, um método retornará índices de probabilidade de que algo ocorra, outro fornecerá dados organizados e compactados, alguns operam como classificadores, outros como reconhecedores de padrões. Por isso entendê-los é inerente para aplicação na resolução de problemas baseados em dados.

Conforme já mencionado, as pesquisas nesta área de conhecimento aplicam estes métodos e considerando a importância das pesquisas e produções científicas no campo da ciência dos dados, assim como dos métodos de análises implícitas empregados para a geração de conhecimento, é importante que também estudos de mapeamento de ciência auxiliem no apontamento das produções científicas.

Neste sentido, o emprego da bibliometria pode resolver importantes questões em torno de informações úteis aos pesquisadores referentes a produção de conhecimento na área de dados dentro da computação.

1.1 MOTIVAÇÕES E OBJETIVOS

Existe vasta produção científica acerca de ciência dos dados, visto sua relevância no cenário de tecnologia atual. E o mapeamento das produções acadêmicas é uma oportunidade de contribuir com valiosas informações a respeito do cenário da produção científica em torno da ciência orientada a dados e dos métodos de análises implícitas.

Os métodos bibliométricos aplicam técnicas estatísticas e quantitativas, a fim de mensurar índices de pesquisas, assim como acompanhar as áreas da ciência

(LOPES Silvia et. al., 2012).

Portanto, entendendo a relevância da ciência orientada a dados e o emprego de métodos de análises implícitas na produção de conhecimento, dentro do escopo da ciência da computação, esta pesquisa tem por objetivo geral mapear por meio da bibliometria a produção científica deste cenário de pesquisa.

Para isso, objetivos específicos precisam ser atendidos, como.

Descrever os conceitos pertinentes a ciência dos dados: história, ciclo de vida e questões de armazenamento, assim como conceituar técnicas de análises implícitas. Também, aplicar análise bibliométrica nas produções científicas abordadas nesta pesquisa.

Ainda, analisar os índices bibliométricos obtidos e também apresentar o panorama das pesquisas científicas referente aos métodos de análises implícitas.

1.2 JUSTIFICATIVA

Mapear a ciência pela perspectiva da bibliometria possibilita identificar as características das pesquisas disponíveis, dando um panorama do atual estado da ciência em determinado tema (SÁNCHEZ; ALBA-RUIZ; RAMIRO, 2014, tradução nossa). O objetivo dessa pesquisa é obter o mapeamento bibliométrico das publicações científicas no campo de ciência da computação que empregam os métodos de análises implícitas de dados.

Ainda, as análises implícitas de dados, representam uma parte significativa das investigações científicas em problemas relacionados a dados (BUFREM et. al., 2016).

Considerando a possibilidade que as publicações ao longo do tempo apresentem características e tendências diferentes das atuais. O mapeamento bibliométrico tende a representar o cenário de pesquisa científica mais atual, podendo então nortear futuras pesquisas (CHUEKE; AMATUCCI, 2015).

A bibliometria é uma técnica baseada em quantificação e estatística com objetivo de mensurar índices referentes a produtividade e abrangência do conhecimento, assim como proporcionar uma visão da ciência, através da

identificação de padrões autorais, publicações e a utilização dos resultados das investigações (LOPES Silvia et. al., 2012).

O ato de realizar uma análise bibliométrica é submeter determinado tema da ciência à uma avaliação das produções científicas em torno de um assunto, este processo ocorre por meio dos diversos indicadores bibliométricos apontando índices quantitativos e qualitativos, de relevância e impacto científico (LOPES Silvia et. al., 2012). Ainda, o levantamento bibliométrico pode ser um método facilitador para criação e gestão de conhecimentos, através do relacionamento dos índices explorados (PEREIRA et. al., 2016).

É entendido que realizar uma análise bibliométrica nas produções científicas em ciência da computação que utilizam os métodos de análises de dados implícitos proporcionará o mapeamento desta área de pesquisa que é intrinsecamente ligada a ciência dos dados. Possibilitando então com o resultado obtido descobrir tendências e informações relevantes para entendimento do cenário de pesquisa e também futuros trabalhos científicos.

A relevância de se realizar um levantamento bibliométrico nesta área de estudo se dá devido aos benefícios em forma de dados advindos unicamente via pesquisa bibliométrica, somado a isso, análises implícitas são amplamente abordadas em produções científicas que buscam solucionar problemas voltados a dados, atuando como técnicas empregadas para classificar, filtrar, processar e identificar padrões entre outras empregabilidades.

1.3 ESTRUTURA DO TRABALHO

Nos capítulos seguintes este trabalho de conclusão de curso está organizado da seguinte forma.

Ciência dos dados é o próximo capítulo, em que é explorado o aspecto histórico do termo remetendo a origem e provável pioneiro desta ciência. Ainda é discorrido em torno das características e relevâncias da área assim como também é exposto considerações sobre o elemento dado e seu ciclo de vida.

Posteriormente é abordado o assunto análises implícitas, este capítulo é

segmentado em duas partes, onde primeiramente de forma introdutória são dissertados conceitos e definições de análises implícitas. Então finalmente na segunda parte do capítulo, na forma de subcapítulos são realizadas exposições de conceitos aplicações e metodologias referentes aos seguintes métodos de análises implícitas: árvores de decisão, classificadores *bayesianos*, redes neurais artificiais, máquinas de vetores de suporte, *k-means*, *k-medoid*, *DBSCAN*, Apriori *e FP-Growth.*

Na sequência são tratados aspectos conceituais referentes a bibliometria, tais como sua história, leis bibliométricas, sua metodologia e contribuições da sua utilização.

Trabalhos correlatos é o próximo item, capítulo em que é agregado e relacionado ao trabalho a experiência e informações contidas em outras pesquisas que estão alinhadas com nosso objetivo de investigação, trazendo de forma adensada em subcapítulos o conteúdo destas publicações.

Por conseguinte, é descrito o processo metodológico utilizado para o desenvolvimento da pesquisa, é apresentada de forma cronológica as etapas realizadas assim como os meios e recursos utilizados.

Em seguida os resultados da pesquisa são apresentados por meio de tabelas e discutidos em relação aos indicadores bibliométricos pertinentes aos dados.

Finalmente como desfecho do trabalho temos a conclusão.

2 CIÊNCIA DOS DADOS

Em pleno quarto paradigma da ciência que é pautado pela produção e consumo de dados, uma área de estudo ganha notoriedade e espaço, sendo relativamente nova e de caráter interdisciplinar, a ciência dos dados por meio das várias disciplinas contidas em si, procura cobrir todos os aspectos pertinentes a dados (CURTY;CERVANTES, 2016).

Em 1962, um estatístico da universidade de *Princeton*, causou um certo desconforto na comunidade acadêmica de estatística e matemática, ao publicar um artigo fazendo referência ao futuro incerto das atividades e pesquisas dos estatísticos (DONOHO, 2017, tradução nossa).

Com o artigo intitulado O Futuro da Análise de Dados *"The Future of Data Analysis",* John Tukey convida seus leitores a aumentar o espectro de observação referente ao processo pelo qual se praticava as análises estatísticas (RALPHS, 2015, tradução nossa).

O autor introduz uma percepção voltada para análise de dados, como sendo algo mais amplo do que os estatísticos da época supunham, apontando para o dado como elemento central das análises (DONOHO, 2017, tradução nossa).

Tukey ainda faz referência a metodologias a serem aplicadas nas análises de dados como: processos para análises de dados, técnicas para interpretação dos resultados, métodos de coletas de dados, formas para melhorar a precisão dos dados além de infraestrutura voltada à análise de dados (DONOHO, 2017, tradução nossa).

Para Tukey o que era visto como inferências estatísticas deveria ser ampliada e redirecionada, a análise de dados não deveria ser uma parte da matemática, mas sim uma nova ciência (STEPHANIE; RAFAEL, 2017, tradução nossa).

A nova ciência sugerida por Tukey, continha quatro principais preocupações que necessitavam esforços:

a) teorias formais da estatística;

b) acelerar o desenvolvimento da computação;

c) o desafio a enfrentar pela quantia de dados gerados nos mais variados campos;

d) a ênfase na quantificação em uma variedade cada vez maior de disciplinas;

As iniciativas de Tukey, não foram imediatamentes reconhecidas e tidas como importantes, no entanto, foram as primeiras evidências de uma nova ciência orientada a dados, sendo em 1962 por John Tukey um prenúncio da ciência dos dados (DONOHO, 2017, tradução nossa).

A ciência dos dados é recente, embora o termo já tenha sido utilizado em 1962 por John Tukey, a sua popularização e uso em grande escala é considerado assunto atual, sua popularização é decorrente do fenômeno da explosão dos dados (AMARAL, 2016).

É conhecida por ser uma área interdisciplinar, visto que é fortemente relacionada com estatística, matemática, conhecimento do negócio e computação (SHCHERBAKOV et. al., 2014, tradução nossa).

Ciência orientada a dados vai além de extrair conhecimento dos dados, considerando que é a ciência que provêm meios e técnicas para que o dado seja tratado em todo o seu ciclo, desde o processo de origem ao descarte(AMARAL, 2016).

Em síntese, a área do conhecimento que é capaz de lidar com as três principais características que envolvem os dados: volume, velocidade e variedade é a ciência dos dados (CURTY;CERVANTES, 2016; SANT'ANA. 2016).

A ciência dos dados lida com os aspectos inerente aos dados desde a produção ao descarte, no entanto, antes de abordarmos o ciclo de vida de um dado, precisamos explanar o que entendemos por dado (AMARAL, 2016).

O dado é a menor parte da informação, possui sentido próprio, porém de forma isolada não caracteriza informação ou conhecimento, portanto, o dado é a matéria-prima da informação (OLIVEIRA, Caroline, 2015).

Além de um conceito global, também pode-se fazer uma distinção dos dados perante seu estado, o dado pode estar em formato eletrônico, analógico ou digital, no entanto, o dado também pode estar em formato não eletrônico que é o

caso de um livro impresso (AMARAL, 2016).

O ciclo de vida do dado é formado pelas seguintes etapas: produção, armazenamento, transformação, análise e descarte, as quais neste capítulo será explanado, no entanto também há autores que incluem mais etapas, como, planejamento, coleta, tratamento, análise, descrição, publicar, descobrir e integrar (AMARAL, 2016; RÜEGG, 2014, tradução nossa).

Existem inúmeras formas pela quais ocorrem a produção dos dados, desde as mais simples como periféricos computacionais, até sensores instalados em diversos dispositivos, um exemplo disso são os *smartphones* (EMC, 2015, tradução nossa).

Dados também podem ser produzidos por transformação, quando um dado sofre alguma transformação seja no formato ou na qualidade, gerando um novo dado (BOTELHO; RAZZONI, 2014).

O processamento de dados, também gera dados, uma hipotética transação bancária cujo processamento gera *logs* é um exemplo de criação por processamento, assim como em uma análise de dados que gera um gráfico, sendo este um tipo de dado (AMARAL, 2016).

Após a criação do dado o mesmo normalmente é armazenado, isto pode ocorrer em memória volátil do dispositivo onde é mantido temporariamente para visualização ou transformação, desta forma se não for posteriormente armazenado em uma unidade de armazenamento não volátil o mesmo será perdido, ou então pode ser gravado em uma unidade não volátil o que permite reaver o dado após desligar o equipamento (WEBER, 2017).

Uma vez armazenado, o dado pode ser replicado por fins de segurança, ou aumento de disponibilidade. Dados podem ser armazenados em diversos formatos de arquivos e bancos de dados. As bases de dados são segmentadas em vários tipos com propósitos distintos (SANT’ANA, 2016).

Bancos de dados dos tipos relacionais, são utilizados para armazenamento de dados que necessitam entre outras coisas manter integridade devido a transações concorrentes, são dados que são relacionados entre tabelas, e organizados em estruturas de linhas e colunas (BOTELHO; RAZZONI, 2014).

Bancos de dados não relacionais, neste caso falando de *Not Only Sql* (NOSQL)*,* são bancos de dados que armazenam estruturas de dados organizados em pares de chave e valor, também há bancos que são orientados a documentos que permitem a criação de estruturas como coleções de arquivos (DONOHO, 2017, tradução nossa).

Referente às tecnologias de armazenamento, as primeiras estruturas eram hierárquicas ou de rede, que foram substituídas pelos bancos de dados relacionais, que ainda são amplamente utilizados, em paralelo aos modelos relacionais foram desenvolvidos os bancos de dados orientados a objetos com intuito de manter uma forte coesão entre modelo de desenvolvimento orientado a objetos e banco de dados, por último a evolução em bases de dados estão em *NOSQL* (KUHNEN, 2016).

Com a variedade dos dados, estruturados, semiestruturados e desestruturados os bancos de dados relacionais mostraram-se não apropriados para o armazenamento e recuperação destas diversas estruturas de dados, sendo então necessário explorar a tecnologia dos bancos de dados não relacionais (AMARAL, 2016).

Transformação de dados é a etapa pela qual os dados passam por algum tipo de processo antes de serem utilizados, esta etapa é fundamental para garantir aspectos como qualidade, confiabilidade e visibilidade dos dados (TAKECIAN, 2014).

Um método clássico de manipular dados é por meio de *Extract Transformation and Load* (*ETL*)*,* é um método amplamente utilizado quando existem diversas fontes de dados e se faz necessário reuni-las em um ponto central a fim de uma análise integrada, ou seja, uma *ETL*, tem a função de extrair dados, transformá-los e carregá-los para outro repositório (BOTELHO; RAZZONI, 2014).

A etapa de análise dos dados é a fase em que ocorre a exploração dos dados contidos em um repositório de dados, este processo ocorre por meio de métodos e ferramentas apropriadas para análise, então como resultado de uma análise: é a extração de *insights,* que são pilares para tomada de decisão, interpretações de cenários, identificação de tendências, validação de hipóteses entre

tantas outras aplicações que podem ser destinadas as informações descobertas neste processo (El ARASS, TIKITO, SOUISSI, 2017, tradução nossa).

A última etapa processual de um dado é o descarte, após passar por um processo de análise e já ter cumprido seu propósito, os dados podem ser então descartados, normalmente este processo é regido por uma governança de dados que indica tempo de armazenamento e condições para o descarte (MOREIRA et. al., 2015, RODRIGUES; SANT'ANA; FERNEDA, 2015).

Não o único, mas o principal objetivo da ciência orientada a dados é transformar dados em informação, e a grande explosão dos dados ofertando dados com volume, velocidade e variedade exigem técnicas adequadas para cumprir este objetivo e por fim agregar valor aos usuários de dados (FAGUNDES; MACEDO; FREUND, 2017).

Para atingir o objetivo macro da ciência dos dados é necessário que os métodos de análises de dados sejam apropriados e, portanto, os métodos de análises implícitas no contexto de ciência dos dados cumprem bem o seu papel visto serem métodos com técnicas apuradas de análise de dados e extração de informação (BUFREM et. al., 2016).

3 ANÁLISES IMPLÍCITAS

Explorar o potencial informativo de um dado ocorre por meio de processos e métodos analíticos, que buscam produzir *insights* a partir de um conjunto de dados analisados, tornando um dado útil, visto que dele se obteve informações (JUNKES, 2014).

O processo de análise de dados varia em detrimento do contexto e tipo de análise pretendida ou necessária, a análise de dados possui várias abordagens e técnicas que se justificam de acordo com o desafio que a coleção de dados implica.

No contexto de ciência dos dados, onde redundância, inconsistência, ruídos, heterogeneidade e volume são aspectos pertinentes a este campo de estudo é fato que métodos específicos de análise de dados sejam empregados para a produção do conhecimento (ROSA, 2018).

Para a descoberta de informações no ambiente de ciência dos dados não são todos os métodos de análises utilizados, mas sim uma determinada categoria, os métodos de análises implícitas de dados (POLA, 2018).

As análises implícitas de dados aplicam-se em casos onde a informação não está explícita e de fácil acesso, em um banco de dados relacional, para fins de exemplo as tabelas, produtos, clientes, e vendas, em que existe um relacionamento explícito entre as entidades mencionadas, a busca por determinadas informações como: quais são os melhores dez clientes, por valores de vendas? (GONÇALVES, 2005).

A resposta é facilmente encontrada com operações de baixa complexidade, neste caso, uma consulta utilizando *Structured Query Language* (SQL) (STEPHANIE; RAFAEL, 2017, tradução nossa).

No entanto, quando a pergunta envolve dados que não estão exatamente estruturados, não possuem relação explícita, possuem fontes e formatos variados e são volumosos, então para estes casos se faz necessário o uso de metodologias mais apuradas, com níveis de complexidade condizente ao cenário de dados. Para estes casos os métodos de análises implícitos são úteis e necessários (AMARAL, 2016).

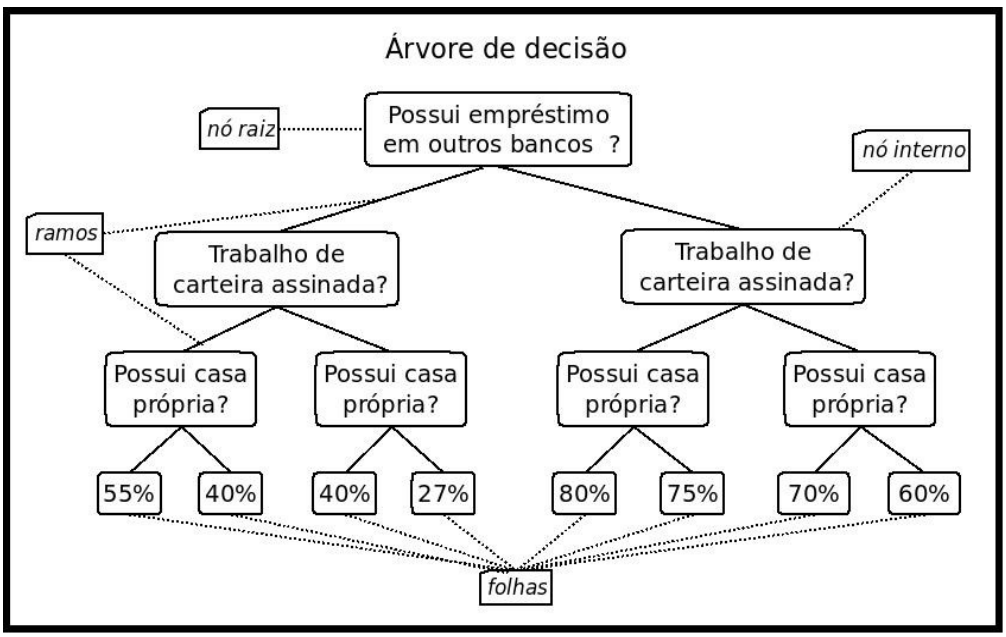
Neste capítulo, de acordo com o autor Fernando Amaral em que na sua obra: Introdução à Ciência de Dados: mineração de dados e big data , aborda as análises implícitas assim como traz exemplos destas técnicas de análise. Sendo assim veremos os principais aspectos de alguns métodos como: árvores de decisão, classificadores *bayesianos*, redes neurais artificiais, máquina de vetores de suporte, *k-means*, *k-medoid*, *Density based spatial clustering of application with noise* (DBSCAN), Apriori e Frequent pattern growth (FP-Growth) (AMARAL, 2016).

3.1 ÁRVORES DE DECISÃO

As árvores de decisão são exemplos estatísticos não paramétricos de aprendizado supervisionado sendo bastante utilizadas em mineração de dados, comumente aplicadas em sistemas de classificação e algoritmos preditivos.

Sua estrutura baseia-se em conceito de ramificação contendo um nó raiz localizado no topo da estrutura, nós internos, folhas e arcos, a partir do nó raiz percorre-se recursivamente aplicando os testes até chegar a classe, (nó folha), que contém a classificação da instância, na figura 1 observa-se a representação de uma árvore de decisão, e sua estrutura (SONG; LU, 2015, tradução nossa).

Figura 1 - Representação de uma árvore de Decisão



Fonte: Própria (2019).

Quanto aos tipos de árvores de decisão, são implementadas duas categorias. Árvores de classificação e as árvores de regressão. A característica que discrimina uma em relação a outra, são as variáveis dependentes, quando as variáveis assumem valores finitos, configura-se uma árvore de classificação, se as variáveis assumem valores contínuos tem-se uma aŕvore de regressão (CHIAVEGATTO FILHO, 2015).

Para que árvores de decisão sejam implementadas e testadas, assim como tenham seu modelo avaliado é necessário que haja um conjunto de dados de treinamento, que é utilizado como entrada para a árvore de decisão, assim como um conjunto de dados de teste, que valida o modelo de dados utilizados (COSTA et. al., 2012).

O ponto inicial de uma árvore de decisão é o nó raiz, que é único em uma árvore, localizado na parte superior e a partir deste percorre-se recursivamente o caminho, aplicando os testes até chegar ao resultado final da classificação contido no nó folha (AMARAL, 2016).

Os nós internos representam os testes aplicados à instância, e o resultado da aplicação dos algoritmos de expressão lógica, (se, então, senão) determinam o caminho da árvore, logo o caminho de uma árvore pertence as regras de classificação. Entre um teste e outro, existe o arco que é um elemento de ligação representando o resultado do teste aplicado ao nó interno anterior (SACHET; SILVA, 2018).

Ao final do percurso da árvore de decisão encontra-se o resultado final, que é a determinação da classe, esta informação é contida nas folhas ou nós terminais, portanto concluindo o percurso da árvore (DAI; JI, 2014, tradução nossa).

A implementação de uma árvore de decisão se dá por meio da aplicação de algoritmos que possuem características convergentes a soluções de árvores de decisão e a escolha do algoritmo está relacionada com os aspectos do problema.

Existem inúmeros algoritmos que implementam árvores de decisão e estes são alguns deles: C4.5, C5.0, *REPTree*, *Random Forest* (RF), *Inductive Decision Tree* (ID3), *Classification And Regression Trees* (CART), *Assistant*, *Chi Square Automatic Interaction Detection* (CHAID), *Induction Algorithm* (CN2), *Scalable Parallel Classifier* (SCALPARC), *Supervised Learning In Quest* (SLIQ), *Scalable PaRallelizable INduction of Decision Tree* (SPRINT), *Trees Parroting Networks* (TREPAN) (DAI; JI, 2014, tradução nossa; ISLER; PITOMBO, 2014).

Portanto, esta sucinta abordagem a respeito de árvores de decisão tem por objetivo contextualizar e situar as árvores de decisão como uma das muitas análises implícitas existentes, assim como os classificadores *bayesianos* que é abordado posteriormente.

3.2 CLASSIFICADORES *BAYESIANOS*

Com o propósito de desenvolver uma metodologia aplicada para a recuperação da informação contida em registros e artigos, M.E. Maron em meados de 1960 por meio de técnicas de indexação de documentos, introduziu o classificador de *bayes*, como um método de classificação automática de documentos, realizando filtragem de conteúdo através da análise de textos, palavras, sentenças e parágrafos (RIEDER, 2019).

Os classificadores *bayesianos* são fundamentados no teorema de B*ayes* que utilizam como base para efetuar a classificação probabilística, sendo meios estatísticos de classificação que categorizam um elemento numa determinada categoria considerando a probabilidade de o objeto pertencer a esta categoria (BONIDIA; BRANCHER; BUSTO, 2018, tradução nossa).

A função de classificar objetos gerando índices de exatidão, quanto a um objeto pertencer ou não a uma classe, são satisfatoriamente desempenhadas pelos classificadores *bayesianos*, visto que apresentam elevadas taxas de exatidão e

apresentam bom desempenho quando aplicados em grandes bases de dados (ZUEGE, 2018).

Ainda, as técnicas de classificação pertencem a área de aprendizado supervisionado, dentro de aprendizado de máquina, visto que, partindo de elementos representantes de classes predefinidas, é capaz de sintetizar em modelos preditivos capazes de indicar a classe de um objeto em consequência dos valores de seus atributos. Estes modelos preditivos são aplicados em várias áreas como biologia, agricultura, medicina e negócios (OLIVEIRA; PEREIRA, 2017).

Além de fundamentar o classificador *bayesiano* é importante analisar o teorema de *bayes,* compreender o funcionamento do teorema que é base para os classificadores *bayesianos* é fundamental, para obter uma compreensão genérica das técnicas decorrentes do teorema de *bayes.*

Os classificadores *bayesianos* baseiam-se no teorema de *Bayes.* Para obter o resultado probabilístico para a classificação, e, no contexto de aprendizado de máquina, isto equivale ao *enésimo* elemento, dada uma nova instância: $A = a_1 \dots a_n$ (1).

$$P(classe|A) = \frac{P(A|classe)\,X\,P(classe)}{P(A)} \quad (2)$$

Como : $A = a_1, a_2 \dots a_n$ (3) *tem-se:*

$$P(classe|a_1 \dots a_n) = \frac{P(a_1 \dots a_n|classe)\,X\,P(classe)}{P(a_1 \dots a_n)} \quad (4)$$

A formula de *bayes* para determinar a classe associada de uma nova instância, calcula-se a probabilidade para todas as classes do escopo, e então rotulando a instância com a classe de maior probabilidade. Seguindo o raciocínio, o denominador da equação é uma constante e, portanto, é anulada da equação chegando a seguinte fórmula simplificada (OLIVEIRA; PEREIRA, 2017).

$$arg\,Max\,P(classe|a_1 \dots a_n) = arg\,Max\,P(a_1 \dots a_n|classe)\,X\,P(classe) \quad (5)$$

Algumas das possíveis aplicações dos algoritmos de classificação é a

identificação de spam e análise de sentimentos.

Por meio de postagens em redes sociais como *twitter* é possível realizar a coleta dos dados e com base nos textos das postagens realizar a classificação e com isso chegar em um resultado, que é a indicação de um possível estado emocional (OLIVEIRA, Cristiano; 2015).

Filtragem de *spam* também é uma aplicação que é dada aos classificadores bayesianos, ao receber um *e-mail*, o classificador com base em modelos de decisão previamente categorizados, consegue filtrar a mensagem classificando-a em algum índice de possibilidade de ser um *spam* (RIEDER, 2019).

Alguns dos exemplos de algoritmos classificadores *bayesianos* são os seguintes: *Tree Augmented Naïve Bayes* (TAN), *using Con-ditional Independence* (CI) Tests, I*nductive Causation Search* (ICS), *General Augmented Naïve Bayes* (GAN), Busca Gulosa (BG) e *Naïve Bayes* (NB) (SÁ; 2014).

Portanto, classificadores *bayesianos* são técnicas probabilisticas derivadas do teorema de *bayes* e são bastante utilizadas em descoberto do conhecimento.

Dando sequência as técnicas de análises implícitas, em seguida são abordados conceitos sobre as redes neurais artificiais.

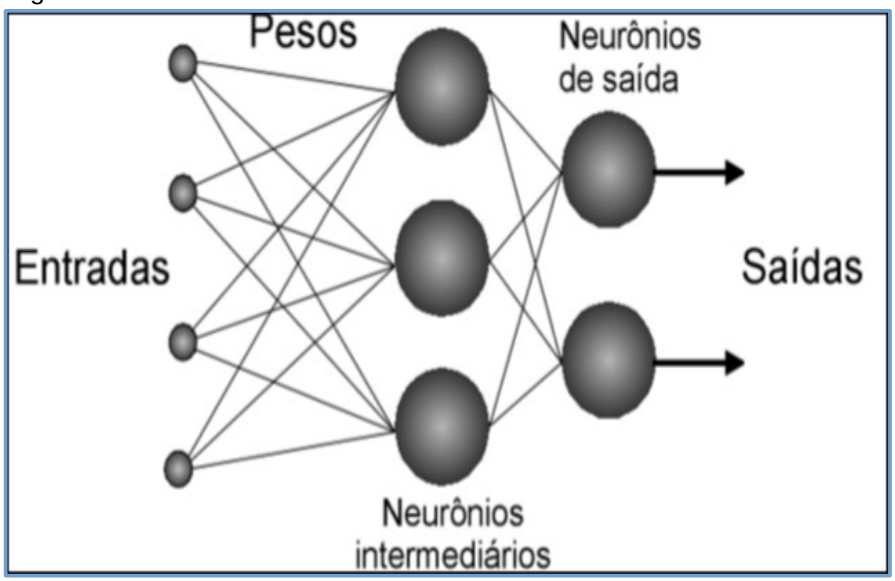
3.3 REDES NEURAIS ARTIFICIAIS

Semelhante à disposição dos neurônios em um cérebro biológico, assim computacionalmente as Redes Neurais Artificiais (RNA) são apresentadas, sendo um conjunto de unidades de processamento onde cada elemento é definido como um neurônio artificial e sobre os dados aplica uma determinada função matemática, são organizados em camadas e conexos entre si, configurando uma rede neural apta a processar, armazenar e transmitir informações imputadas (BINOTI D; LEITE; BINOTI M, 2014).

A neurôdinamica e a arquitetura de uma rede variam em função de propagação da informação e alimentação da rede, basicamente a arquitetura é semelhante ao apresentado na figura 2, as entradas recebem as informações e de

acordo com o conhecimento armazenado na rede, na figura nomeado como peso, é realizada a ligação com a camada de processamento, então os neorônios intermediários estabelecem conexão com os neurônios de saída, configurando o resultado da rede (FÁVERO; ZOUCAS, 2016).

Figura 2 - Redes Neurais Artificiais



Fonte: MARQUES;ULSON (2018).

Redes neurais artificiais são uma classe de algoritmos não lineares pertencentes ao campo da inteligência artificial (IA), que dentre outros propósitos são bastante utilizadas em descoberta de conhecimento, sendo um recurso na metodologia de mineração de dados, onde atuam na predição de eventos como por exemplo, antever a possibilidade de uma transação bancária ser fraudulenta (SILVA et. al., 2013).

Este importante segmento da IA, possui aspectos tal como, capacidade de generalização, adaptação, correlação e aprendizado. Sendo relevantes para determinadas tarefas como: previsão e classificação, logo várias áreas de conhecimento aplicam RNAs em suas análises (CUNHA, 2018).

Diante do cenário de *big data* uma característica relevante das RNAs se faz proveitosa. O aspecto de que as RNAs possuem capacidade de trabalhar com dados imprecisos e complicados, o que é abundante no cenário de ciência dos dados, e destes extrair padrões e detectar tendências. Somado a isso a redução de

custo de aquisição e o aumento da capacidade de processamento fomentam o desenvolvimento de RNA no emprego de descoberta e correlação de conhecimento (MARQUES;ULSON, 2018).

A inteligência artificial, disponibiliza inúmeras arquiteturas de redes, onde cada arquitetura lida com determinados tipos de problemas, onde seja mais apropriado uma ou outra arquitetura de rede neural, sendo assim segue alguns exemplos de modelos de redes neurais disponíveis: *Deep Neural Networks* (DNN), *Convolutional Neural Network* (CNN), Network *Multilayers Perceptrons* (MPNN), *Recurrents Neural Network* (RNR), *Long Short-Term Memory* (LSTM), *Hopfield Networks,* Boltzmann machines, *Deep Belief Network* (DBN), *Deep Auto-Encoders* (DAE), *General Adversarial Network* (GAN) e *Deep Neural Network Capsules* (CapesNet).

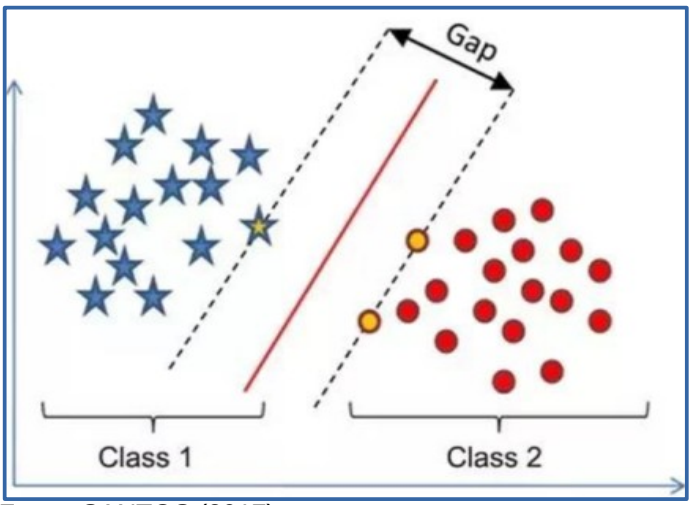
Este subcapítulo conceitua as RNA's de forma genérica, não cabendo aqui conceituá-las individualmente, mas é importante que se tenha o conhecimento da existência da vasta disponibilidade de recursos em RNA (FAJAR; RACHMAD, 2018, tradução nossa; FÁVERO; ZOUCAS, 2016 ; HAN; MAO; DALLY, 2016, tradução nossa; MARQUES;ULSON, 2018; NELSON, 2017).

3.4 MÁQUINA DE VETORES DE SUPORTE

As Máquinas de Vetores de Suporte (SVM), são técnicas de aprendizado de máquina do tipo classificador linear binário não probabilístico, que foram introduzidas por Vladmir Vapnik em 1995. Esta técnica fundamenta-se na teoria de aprendizado estatístico como estratégia de classificação, os principais aspectos deste método são excelente capacidade de generalização e capacidade para lidar com grandes quantidades de dados (DOSCIATTI; PATERNO; PARAISO, 2013).

Este classificador em seu modo mais simples, o método linear, estabelece um hiperplano, um espaço com uma margem segregando um grupo de elementos positivos e negativos, binários em um espaço com grande dimensão, conforme é ilustrado na figura 3. (MARQUES; ULSON, 2018).

Figura 3 - Método de Máquina de Vetor de Suporte.



Fonte: SANTOS (2017).

Considerando que podem existir tantas quantas possíveis alternativas para o lado que faça a divisão desses objetos, a finalidade do método é expandir ao máximo o espaçamento de um lado aos objetos negativo e positivo mais próximos.

Separar linearmente um *dataset,* não é sempre possível, portanto não existe um hiperplano que separe todos os pontos negativos e positivos. Nesse circunstância é aplicada uma correção a um exemplo que falhe em se posicionar no seu lado correto.

Também, SVM's podem ser utilizadas como classificadores não lineares. Sendo utilizadas funções *kernel* que adicionam não-linearidades *gaussianas* e polinomiais (SANTOS; 2017).

Contudo o método disponibiliza uma solução aplicável em situações binárias, essa tratativa pode ser ampliada para problemas com múltiplas classes. Tal resultado pode ser atingido por meio de duas estratégias: executando o teste de uma categoria contra todas as outras ou uma tratativa com inúmeros testes entre duas categorias (AHMED et. al., 2012, tradução nossa).

Para o treinamento de uma SVM necessita-se solucionar a limitação de otimização de programação quadrática. Em 1998, John Platt recomenda o algoritmo *Sequential Minimal Optimization* (SMO), que possibilita uma solução para quebrar o problema de programação quadrática original em uma série de problemas menores.

Esta estratégia possibilitou treinar grandes conjuntos de testes, visto a redução do tempo de treinamento (SANTOS; 2017).

Portanto as SVM's também fazem parte das técnicas de análises implícitas, sendo um tipo de classificador assim como *k-means,* próxima técnica a ser abordada.

3.5 *K-MEANS*

A ciência dos dados por meio de suas várias disciplinas disponibiliza inúmeros algoritmos para se lidar com os dados, e estes procedimentos computacionais de acordo com sua finalidade e colocação no campo dos dados são classificados em categorias. Entre estas categorias existem os algoritmos de *clustering*, cuja finalidade é realizar agrupamentos de itens de forma a seguir um padrão, e um dos vários algoritmos do tipo *cluster* é o *k-means* (OLIVEIRA; 2018).

Em análise de dados, uma técnica de *clusterização* pode ser aplicada ao conjunto de dados visando obter informações e padrões referente a estrutura de dados que está sendo investigada, em análise exploratória de dados, não se tem informações acerca dos dados, nestes casos a funcionalidade de agrupamentos de dados por meio de *k-means* é uma ferramenta valiosa (MORISSETTE; CHARTIER, 2013, tradução nossa).

Introduzido por James B. MacQueen em 1967, *k-means* é um dos métodos hierárquicos mais populares de *clusterização*, seu processo de agrupamento é fundamentado em alocação de objetos, considerando a média do cluster mais próximo (ADAZIMA; BUSTAMAN; ALDILA, 2019, tradução nossa).

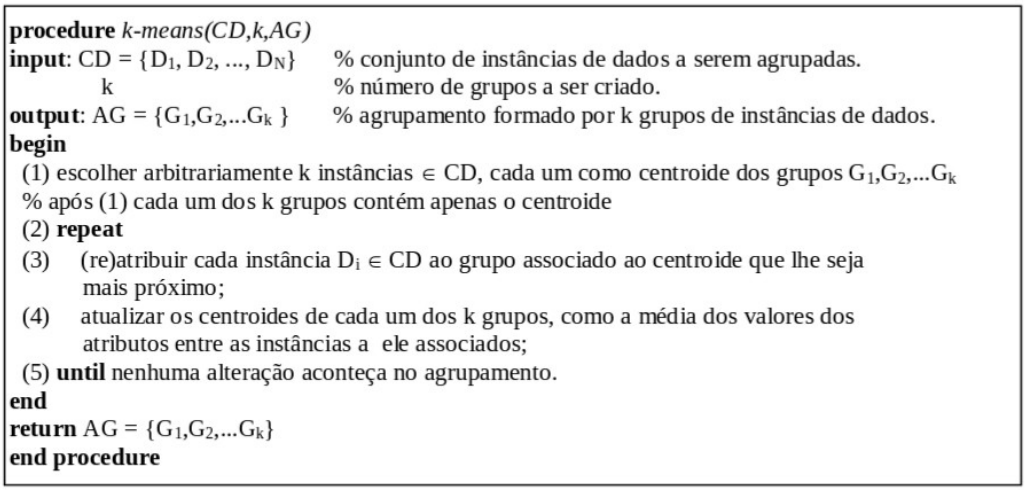
O método de clusterização *k-means* pode ser usado de forma exploratória para descobrir grupos significativos dentro de um conjunto de dados, ou pode servir como ponto de partida para análises mais avançadas. Assim, aplicações de *clustering* são abundantes em aprendizado de máquina e análise de dados, incluindo, análise de expressão genética, segmentação de mercado, análise de redes sociais, segmentação de imagens detecção de anomalias entre outros (WANG; GITTENS; MAHONEY, 2019, tradução nossa).

Quanto aos algoritmos de clusterização do tipo *k-means* destacam-se três deles, devido serem os mais utilizados como técnica de clusterização baseados em k-means, algoritmo *ForgyLloyd*, algoritmo *MacQueen* e o algoritmo *Hartigan* e *Wong*.

Não há um algoritmo melhor de forma absoluta, a escolha do algoritmo está associada com as características do *dataset,* tamanho e número de variáveis, alguns autores sugerem treinar o *dataset* com vários algoritmos para entender a base de dados (MORISSETTE; CHARTIER, 2013, tradução nossa).

Na figura 4 é disponibilizado um pseudocódigo, comentado do algortimo *k-means,* sua entrada, processamento, iterações e sáida.

Figura 4 – Pseudocódigo comentado *k-means*



Fonte OLIVEIRA (2018).

Os algoritmos *k-means* são computacionalmente onerosos, e este custo está relacionado ao número de *clusters* contidos no exemplo, uma escolha visando o menor custo computacional implicaria em testar todos os centróides, que são os vetores de cada *cluster* o que elevaria o problema para um caso NP-Difícil.

No entanto, de forma heurística inicia-se os centróides, e em seguida modificam-se a partição para minimizar a soma das distâncias de cada objeto para com a média do cluster ao qual o objeto pertence.

O objetivo disto é assinalar cada objeto com a partição cuja média (centróide) está mais próxima do objeto. Esse passo gera uma nova partição em que a soma das distâncias tende a ser menor do que no particionamento anterior. O passo de redução da soma das distâncias é repetido até que essa melhora seja muito pequena ou nula.

Se a redução levar a menos do que *k* partições, uma das partições (geralmente a que possui a maior soma de distâncias da média) é dividida em duas ou mais, até atingir o número de k partições desejadas pelo usuário (MELO; 2005).

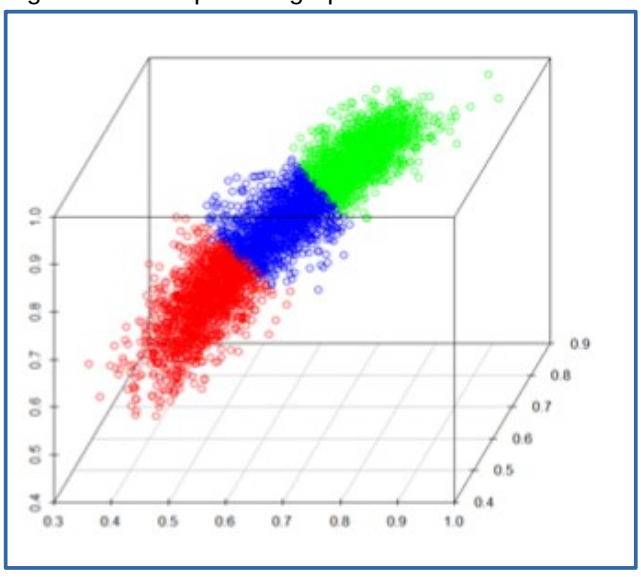
Portanto o fundamento do método *k-means* é minimizar a variância no *cluster*, isto é, os objetos pertencentes a um *cluster* precisam estar o mais próximo possível do centro do *cluster*.

Então, *k-means* é um método baseado em centróide, por outro lado o *k-medoid* que também é uma algoritmo de classificação, baseia-se no *medoid,* este algoritmo é tratado em seguida.

3.6 *K-MEDOID*

A *clusterização* é uma atividade relacionada ao aprendizado de máquina não supervisionado, aplica-se esta técnica sobre os conjuntos de dados de forma a obter conjuntos disjuntos, separando-os em *clusters,* então o nome *clusterização*, resultando em conjuntos de dados onde os elementos similares formam grupos semelhantes ao ilustrado na figura 5 (KAUR N; KAUR U; SINGH, 2014, tradução nossa).

Figura 5 - Exemplo de agrupamento



Fonte: BRITO J; SEMAAN; BRITO L (2014).

Nas técnicas de particionamento de dados assim como o *k-means*, o *k-medoid* também faz parte desta categoria de algoritmos, sendo um procedimento de *clusterização*.

K-medoid é um dos algoritmos mais populares de *clusterização* hierárquica. Ao contrário do algoritmo *k-means* que utiliza os centróides como estimativa para o *cluster* o *k-medoid* devido as dissimilaridades arbitrárias utiliza o *medoid* (SCHUBERT; ROUSSEEUW, 2018, tradução nossa).

No domínio de entendimento de *clusters* define-se que o objeto com a menor dissimilaridade (diferença) do *cluster* (demais objetos) é conhecido como *medoid* e passa a ser referência para o algoritmo de *k-medoid* (BRITO J; SEMAAN; BRITO L, 2014).

Um exemplo de pseudocódigo comentado, descrendo os passos de execução do algoritmo k-medoid é demonstrado na figura 6.

Figura 6 – Pseudocódigo comentado *k-medoid*

```
/*
   K número de cluster para executar
*/

k-medoid(k)

/*
   seleciona k pontos do objeto representativo inicial
   do k cluster de entrada
*/

iniciliza() ;

repete():
        /*
          Atribui cada ponto do cluster com medoid mais próximo (m)
          Randômicamente  seleciona um objeto(i) não representativo
          Soma o custo total de troca  'S' do medoid com objeto(i)
        */

        Se S < 0 :

                /*
                  Troca 'm' com objeto(i) para executar novo conjunto de medoids.
                */

Para();
/*
   Quando o critério o atingido.
*/
```

Fonte: Autor (2019)

Muito embora seja um dos algoritmos mais conhecidos e utilizados no segmento de clusterização, o *k-medoid* possui um reconhecido problema relacionado a performance, isto causa a elevação do tempo de execução (KAUR N; KAUR U; SINGH, 2014, tradução nossa).

Na tentativa de amenizar este problema algumas abordagens referentes a este algoritmo surgiram com possíveis melhorias ao problema de agrupamento, estes algoritmos são: *Biased Random-Key Genetic Algorithm* (BRKGA), algoritmo genético de chaves aleatórias viciadas, *Partitioning Around Medoids* (PAM), particionamento acerca de medoids, *Clustering Large Applications* (CLARA), *clusterização* em aplicações grandes e *Clustering Large Applications based upon RANdomized Search* (CLARANS) *clusterização* em aplicações grandes baseado e pesquisas randômicas (SCHUBERT; ROUSSEEUW, 2018, tradução nossa).

Estes algoritmos, são algumas das técnicas que buscam melhorar a performance de *clusterização* do *k-medoid*.

A metodologia básica de execução do *k-medoid* é baseada no princípio de minimizar a soma das diferenças entre cada elemento e seu ponto de referência correspondente, sendo assim a base do método *k-medoid*, então encontra-se o *k cluster* do objeto de forma arbitrária, sendo o *k cluster* a representação do *cluster*, determinando então o medoid (KAUR N; KAUR U; SINGH, 2014, tradução nossa).

Dessa forma este algoritmo é mais um exemplo de técnica de análise implícita, neste caso voltada a agrupamentos e *clusterizações*, ainda na linha de *cluster* em seguida á abordado um algoritmo endereçado a dados espaciais, o *DBSCAN*.

3.7 *DBSCAN*

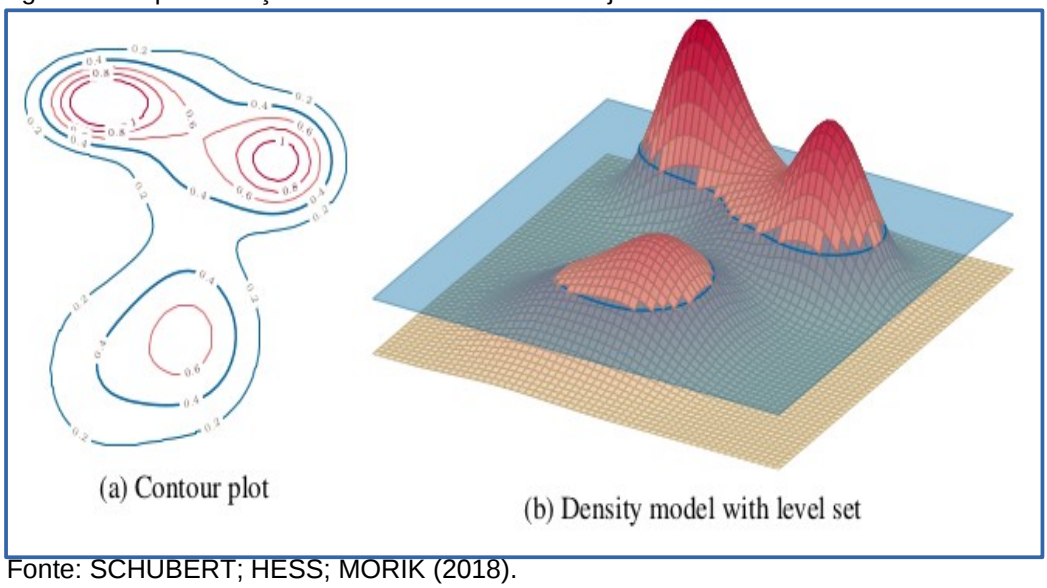
Em 1996 os algoritmos de *clusterização*, ainda não apresentavam boas soluções para bancos de dados espaciais, foi então que por meio do artigo *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise* dos autores Martin Ester, Hans-Peter Kriegel, Jiirg Sander e Xiaowei Xu, o

algoritmo *DBSCAN* foi apresentado pela primeira vez (REHMAN; ASGHAR; FONG, 2014, tradução nossa).

DBSCAN assim como *k-means* e *k-medoid* estão contidos no grupo dos algoritmos de clusterização utilizados em aprendizado de máquina, *DBSCAN,* é acrônimo para D*ensity-Based Spatial Clustering of Applications With Noise,* agrupamento espacial baseado em densidade de aplicativos com ruído, é uma importante técnica de *clusterização* em bases de dados espaciais por densidade (HE et. al., 2013, tradução nossa).

Este algoritmo é especialista em lidar com *clusters* quais têm como característica, haver áreas densas e isoladas de dados, assemelhando-se com ilhas, conforme figura 7, estes *clusters* são conhecidos como *cluster density-based.* Ainda *DBSCAN* é uma técnica para encontrar estes tipos de *clusters* em bancos de dados (SCHUBERT; HESS; MORIK, 2018, tradução nossa).

Figura 7 - Representação de densidade em um conjunto



Fonte: SCHUBERT; HESS; MORIK (2018).

Também, o que diferencia o algoritmo *DBSCAN* dos demais algoritmos de clusterização é o método base pelo qual é realizado o agrupamento, os demais algoritmos fundamentam-se na distância entre os objetos enquanto que o *DBSCAN* consiste na densidade do *cluster*, e os métodos que seguem esta vertente são

conhecidos como métodos de *clusterização* baseado em densidade (CASSIANO; PESSANHA, 2014).

Na figura 7 é apresentado um pseudocódigo do funcionamento do algoritmo *DBSCAN.*

Nestes métodos os *clusters* são as regiões mais densas, separadas por regiões de menor densidade, conhecidas por ruídos, a capacidade de identificar este tipo característico de *cluster*, torna o *DBSCAN* especialista neste segmento (CASSIANO; PESSANHA, 2014).

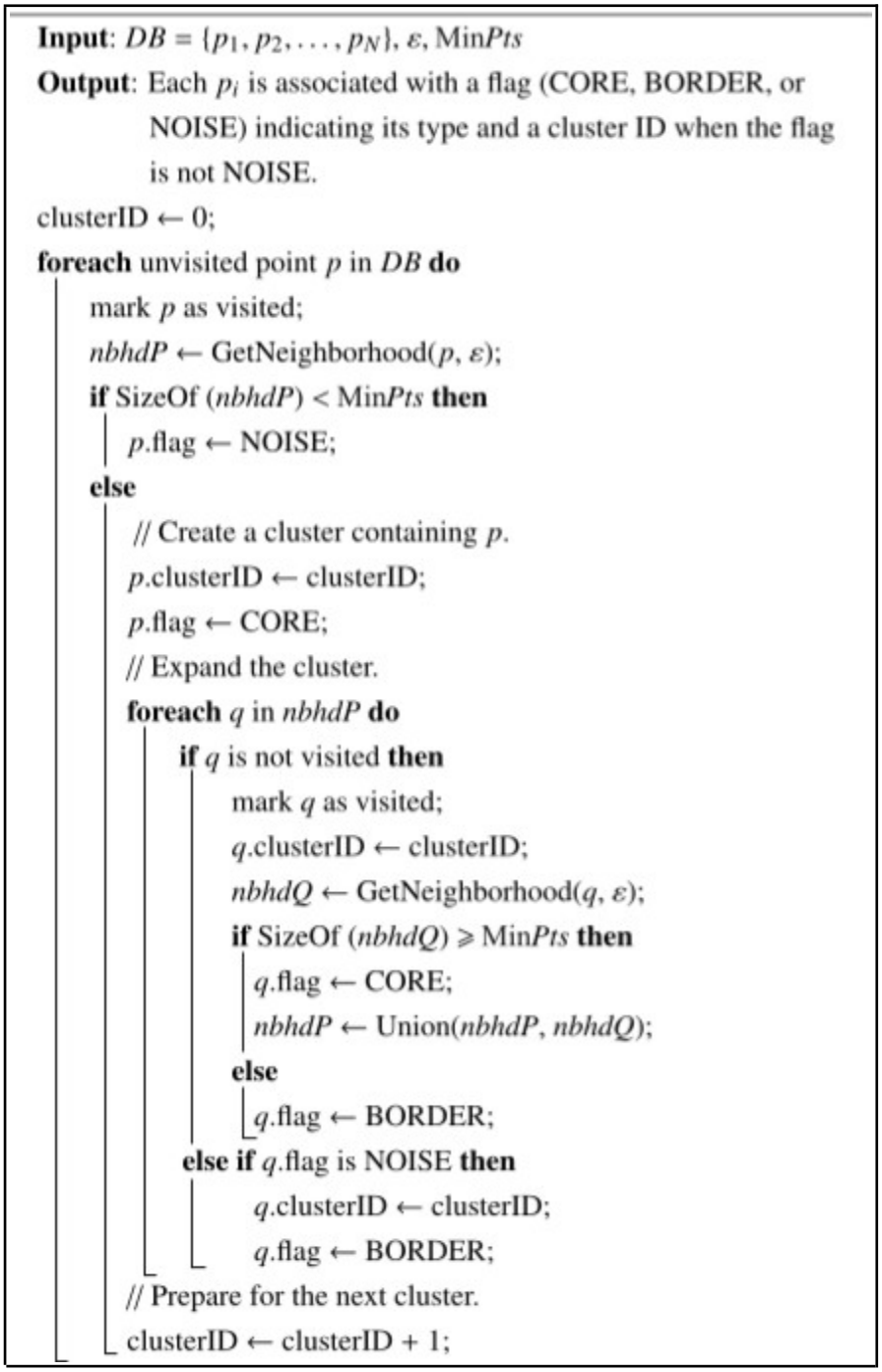
Esta técnica de *clusterização* é amplamente aplicada em grandes *datasets,* porém, apresenta três grandes problemas relacionados a processamento paralelo. Estes problemas são relacionados a balanceamento de carga em processos paralelos, escalabilidade devido a rotinas críticas não serem paralelizáveis e por não serem designados para compartilhamento de ambientes o que os torna menos portáveis em paradigmas de processos paralelos (HE et. al., 2013, tradução nossa).

Na tentativa de resolver estes problemas muitas vertentes deste algoritmo foram desenvolvidas, pesquisadores a partir do algoritmo padrão *DBSCAN* tentaram desenvolver versões melhoradas do algoritmo. Estes novos algoritmos são *VDBSCAN*, *FDBSCAN*, *DD_DBSCAN*, *IDBSCAN*, *GRIDBSCAN* e *EDBSCAN* (REHMAN; ASGHAR; FONG, 2014, tradução nossa).

Ao passo que é informado outras variantes do algoritmo *DBSCAN,* não é realizado o aprofundamento em detalhes de cada uma delas, ora, o objetivo é contextualizar e informar a existência e possível uso das mesmas, até por que não cabe aqui ampla discussão, mas sim, apontar conteúdos fundamentais a respeito de técnicas algorítmicas compatíveis com o conceito de análise implícita.

Na figura 8 é apresentado um pseudocódigo da lógica empregada no algoritmo DBSCAN.

Figura 8 – Pseudocódigo comentado do algoritmo *DBSCAN*



```
Input: DB = {p_1, p_2, ..., p_N}, ε, MinPts
Output: Each p_i is associated with a flag (CORE, BORDER, or
        NOISE) indicating its type and a cluster ID when the flag
        is not NOISE.
clusterID ← 0;
foreach unvisited point p in DB do
    mark p as visited;
    nbhdP ← GetNeighborhood(p, ε);
    if SizeOf (nbhdP) < MinPts then
        p.flag ← NOISE;
    else
        // Create a cluster containing p.
        p.clusterID ← clusterID;
        p.flag ← CORE;
        // Expand the cluster.
        foreach q in nbhdP do
            if q is not visited then
                mark q as visited;
                q.clusterID ← clusterID;
                nbhdQ ← GetNeighborhood(q, ε);
                if SizeOf (nbhdQ) ⩾ MinPts then
                    q.flag ← CORE;
                    nbhdP ← Union(nbhdP, nbhdQ);
                else
                    q.flag ← BORDER;
            else if q.flag is NOISE then
                    q.clusterID ← clusterID;
                    q.flag ← BORDER;
        // Prepare for the next cluster.
        clusterID ← clusterID + 1;
```

Fonte: HE et. al., (2013)

Por conseguinte é abordado uma importante técnica ferramental para análises focadas em descoberta do conhecimento, Apriori.

3.8 APRIORI

Apriori é um algoritmo utilizado em mineração de dados para associação de conhecimento, proposto em 1994 por Agrawal e Srikant. Baseia-se previamente em propriedades de conjuntos de informações, tais como, *log* de transações, *datasets* e banco de dados, sendo estes um *input* necessário para o algoritmo, o nome Apriori se justifica, justamente porque o algoritmo necessita de uma fonte de conhecimento prévio (BONIDIA; BRANCHER; BUSTO, 2018 ; ELGAML et. al., 2015, tradução nossa; SIDHU et. al., 2014, tradução nossa).

Assim como outros algoritmos associativos empregados em mineração de dados, o algoritmo Apriori também usa o método iterativo para desempenhar sua funcionalidade, o que é uma característica e também um problema, visto que realizar processos iterativos em grandes bases de dados de forma que toda a base de dados seja rastreada torna-se bastante dispendioso de recursos computacionais.

Um importante fundamento empregado neste algoritmo se determina pela seguinte regra, para todo padrão de comprimento *k,* que não for frequente no conjunto de dados analisados, logo o comprimento *k+1* também não o será, e para todo padrão *k* que for frequente o padrão *k-1* também será, está característica é determinada pela propriedade *anti-monotonicity* (CASTRO; 2014).

Então, por meio de processo iterativo gera-se conjuntos de padrões candidatos de comprimento *k+1* a partir de um conjunto de padrões de frequência de comprimento *k,* onde $k \geq 1$ e portanto é verificado a frequência de ocorrência na base de dados, na figura 9 a lógica empregada neste algoritmo é demonstrada por meio de pseudocódigo comentado (NANDI et. al., 2015).

Figura 9 – Pseudocódigo comentado algoritmo Apriori

```
/* Parametros de entrada:
  - base de dados D e valor de suporte mínimo: min_sup */

/* Saída
  Conjunto L de todos os itemsets frequentes. */

Função apriori_main (D, min_sup)
   L1 = {conjunto dos itemsets frequentes de tamanho 1 contidos em D};
   para ( k = 2; L k-1 ≠ 0; k++)
            Ck = apriori-gen (L k-1);
            para todas transações t Є D fazer
               Ct = subset (Ck+ t);
               para todos candidatos c Є Ct fazer
                     c.count ++;
               fim para
            fim para
```

Fonte: Autor (2019).

Ainda, fundamentando o método Apriori, a existência de um conjunto de elementos implica na existência de um subconjunto destes elementos, sendo necessário o algoritmo realizar inúmeras análises até que encontre os dados transacionais que ocorrem com maior frequência, resultando deste processo um conjunto de elementos dito candidatos (SACHET; SILVA, 2018).

O método recursivo de escaneamento ao banco de dados para a busca de conjuntos candidatos, é uma das três etapas que compõem o processo que este algoritmo executa, as fases são, geração dos conjuntos candidatos, poda dos conjuntos candidatos e cálculo do suporte (BONIDIA; BRANCHER; BUSTO, 2018).

A geração dos conjuntos candidatos já fora explicada anteriormente, cabendo então explicar as demais fases envolvidas no processo do algoritmo Apriori.

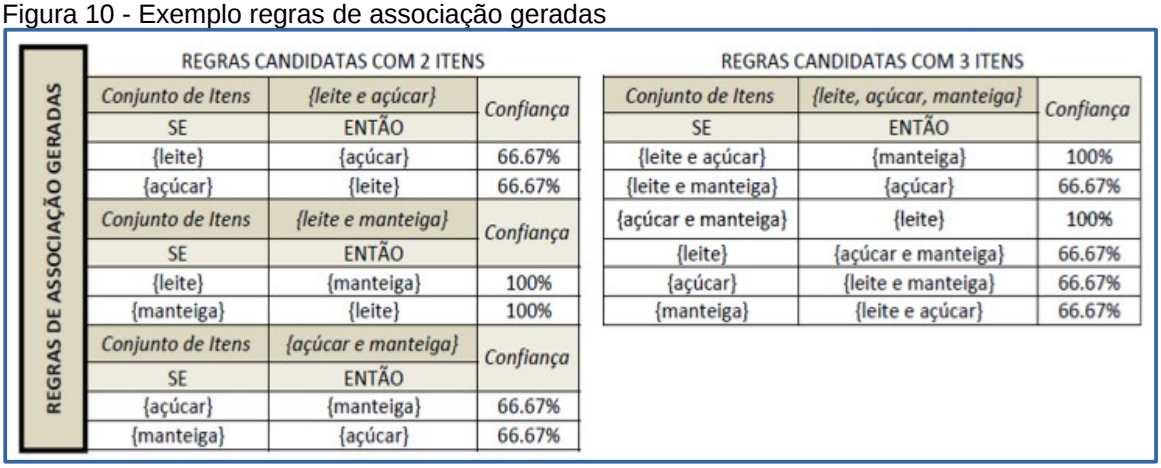
A etapa de poda de conjunto ou também conhecida como *prune* consiste em eliminar os conjuntos que não são frequentes, em suma, é realizado um ajuste no conjunto de elementos (VASCONCELOS; CARVALHO, 2018).

Quanto ao cálculo de suporte do candidato, basicamente para cada transação que for frequente, isto é, existir, é incrementado então no contador do

candidato se este é um ponto de corte, considerando o nível de suporte mínimo definido (CASTRO; 2014).

Após todas as etapas do algoritmo de geração de conjuntos frequentes serem concluídas é então realizada a extração das regras de associação que basicamente consiste em aplicar um índice mínimo de confiança, para cada subconjunto dos conjuntos frequentes resultando na extração das regras que satisfazem o índice mínimo de confiança conforme esta exemplificado na figura 10, os conjuntos de itens, as regras candidatas e o percentual de confiança para a regra. (CASTRO; 2014).

Figura 10 - Exemplo regras de associação geradas



REGRAS DE ASSOCIAÇÃO GERADAS

REGRAS CANDIDATAS COM 2 ITENS

Conjunto de Itens	*{leite e açúcar}*	*Confiança*
SE	ENTÃO	
{leite}	{açúcar}	66.67%
{açúcar}	{leite}	66.67%
Conjunto de Itens	*{leite e manteiga}*	*Confiança*
SE	ENTÃO	
{leite}	{manteiga}	100%
{manteiga}	{leite}	100%
Conjunto de Itens	*{açúcar e manteiga}*	*Confiança*
SE	ENTÃO	
{açúcar}	{manteiga}	66.67%
{manteiga}	{açúcar}	66.67%

REGRAS CANDIDATAS COM 3 ITENS

Conjunto de Itens	*{leite, açúcar, manteiga}*	*Confiança*
SE	ENTÃO	
{leite e açúcar}	{manteiga}	100%
{leite e manteiga}	{açúcar}	66.67%
{açúcar e manteiga}	{leite}	100%
{leite}	{açúcar e manteiga}	66.67%
{açúcar}	{leite e manteiga}	66.67%
{manteiga}	{leite e açúcar}	66.67%

Fonte: CASTRO (2014).

Apriori é uma das técnicas de análises implícitas bastante utililizadas em mineração de dados, assim como a próxima técnica que será abordada, o algoritmo *FP-Growth.*

3.9 *FP-GROWTH*

O algoritmo *FP-Growth* empregado em mineração de dados é utilizado com o objetivo de identificar padrões de transação em bancos de dados, ou seja, detectar a frequência padrão, *Frequent Pattern* (FP) com que transações ocorreram, tornando evidente as transações que satisfaçam um determinado índice de frequência preestabelecido (IKHWAN et. al., 2018, tradução nossa).

Este algoritmo foi introduzido no ano 2000 por Jiawei Han, Jian Pei, Yiwen Yin atráves do artigo *Mining Frequent Patterns without Candidate Generation.* A proposta do algoritmo é obter o mesmo resultado em termos de pesquisa dos dados, porém sem o custo computacional da etapa *generation candidate,* vista como onerosa no algoritmo Apriori (DIZON et. al., 2019, tradução nossa).

Uma característica deste método é o emprego da estratégia de dividir para conquistar e diferentemente dos demais algoritmos de mesma finalidade, o *FP-Growth*, não utiliza a etapa de geração de candidatos, mas sim, armazena um conjunto de dados de forma compacta em uma estrutura de árvore conhecida como *Frequent Pattern tree (FP-Tree)* da qual realiza a extração dos itens frequentes, esta abordagem proporciona melhor desempenho se comparado com outros métodos (NANDI et. al., 2015).

FP-Tree é uma estrutura de dados densa baseada em árvore, em que são armazenados de forma comprimida as transações do banco de dados analisado. Em posterior análise, o algoritmo *FP-Growth* recorre a *FP-Tree* extraindo diretamente deste os conjuntos de dados frequentes. O fato de não fazer recorrentes escaneamentos em toda a base de dados torna este algoritmo mais performático (IKHWAN et. al., 2018, tradução nossa).

O método de associação *FP-Growth* é separado nas seguintes fases, criação da estrutura de árvore *FP-Tree*, divisão da *FP-Tree e* mineração das estruturas menores.

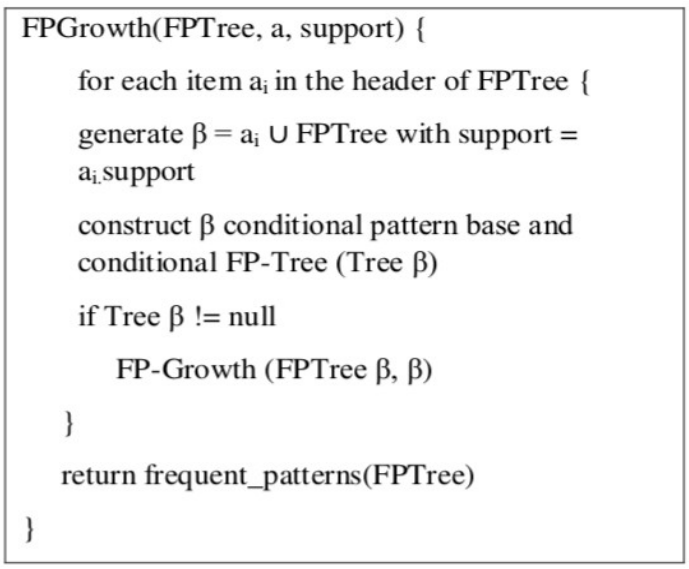
A primeira fase do algoritmo é a criação da estrutura de árvore *FP-Tree*, neste passo o método inicia compactando a informação oriunda do *dataset* em questão, e nesta estrutura ficam representados os conjuntos frequentes (ELGAML et. al., 2015, tradução nossa).

Após montar a árvore *FP-Tree* com os conjuntos frequentes o algoritmo realiza a divisão da árvore em várias unidades de bancos de dados onde cada banco contêm as informações de cada padrão frequente.

Finalmente, cada um dos bancos de dados resultante da divisão do FP-Tree é minerado separadamente (CASTRO Rui, 2016).

Conforme descrição dos passos lógicos do algoritmo *FP-Growth,* nos parágrafos superiores. A figura 11 representa a execução destes processos algorítmicos.

Figura – 11 Pseudocódigo do algoritmo FP-Growth



```
FPGrowth(FPTree, a, support) {
    for each item a_i in the header of FPTree {
    generate β = a_i ∪ FPTree with support =
    a_i.support
    construct β conditional pattern base and
    conditional FP-Tree (Tree β)
    if Tree β != null
        FP-Growth (FPTree β, β)
  }
  return frequent_patterns(FPTree)
}
```

Fonte : DIZON et. al., (2019).

Ainda, sobre as fases do algoritmo, a primeira fase é considerada a mais importante, visto que o algoritmo realiza escaneamento completo da base de dados por duas vezes.

No primeiro escaneamento, o propósito é realizar a busca pelas transações frequentes e atribuir uma contagem para cada transação, através desta contagem organiza uma ordenação ascendente deixando *null* como nó raiz.

No segundo escaneamento é construída a estrutura de árvore, baseando-se nos arranjos previamente feitos no primeiro escaneamento (ELGAML et. al., 2015, tradução nossa).

Este capítulo, é um embasamento teórico dos fundamentos de análises implícitas e por consequência das análises específicas que aqui discorremos. A pesquisa bibliométrica investigará trabalhos que empregaram estas técnicas, logo a fundamentação teórica é desejável, para contextualização e referencial das análises.

4 BIBLIOMETRIA

A geração de informação é um fenômeno em crescimento acelerado e a produção de conteúdo científico acompanha este movimento, logo, cabe a comunidade científica entender os aspectos e características do conhecimento produzido acerca das diversas áreas da ciência. Para que isto ocorra é necessário mensurar a produção do conhecimento por meio de investigações específicas para esta finalidade, como é o caso das pesquisas bibliométricas (ROQUE, 2012).

A bibliometria é uma área de estudos do campo da ciência da informação, que de forma quantitativa e estatística avalia as publicações científicas e decorrente disso gera indicadores relativos à produção de conhecimento da comunidade científica. O resultado, os indicadores, servem como elementos de referência para novas publicações, direcionamentos de pesquisas, acompanhamento das diversas áreas do conhecimento, produção e disseminação do conhecimento, comportamento de autoria, disponibilização e uso dos resultados das pesquisas (MEDEIROS; VITORIANO, 2015; SÁNCHEZ; ALBA-RUIZ; RAMIRO, 2014, tradução nossa).

Quanto ao surgimento da bibliometria, dependendo do ponto investigado, podendo ser o termo propriamente dito, a área de estudo ou a prática de medir bibliografias, as pesquisas apontarão origens distintas.

A revisão bibliográfica aponta que o primeiro método aplicado com intuito de mensurar obras literárias, e, entende-se que em virtude da época também científicas, foi a esticometria, procedimento empregado pelos antigos gregos que consistia em mensurar a extensão dos manuscritos, estes eram medidos em esticos que correspondiam a quinze ou dezesseis sílabas contidas em trinta e cinco ou trinta e seis letras, a cada obra era atribuído sua medida equivalente, e esta informação era permanente, sendo utilizado com as finalidades de índice de cálculo para pagamento ao copista, evitar supressões e interpolações textuais (ALVARADO, 2007).

Já em 1869 Francis Galton realizou um experimento objetivando identificar cientistas eminentes através das menções de nomes em bibliografias

selecionadas, no fim do século XIX, Campell investigou a dispersão de assuntos em publicações empregando métodos estatísticos (MEDEIROS; VITORIANO, 2015).

Alguns pesquisadores atribuem o surgimento da bibliometria a Cole e Eales quando estes em 1917 publicaram uma análise estatística da história da disciplina de anatomia comparada (LOPES Silvia et. al., 2012).

No entanto a bibliometria teria inicialmente sido difundida como "*bibliografia estatística*", por E. Wyndham Hulme no ano de 1923, e a nomenclatura como é conhecida hoje teria sido criada em 1934 por Paul Otlet em sua obra "*Traité de Documentacion*" (MEDEIROS; VITORIANO, 2015).

Entretanto credita-se a Pritchard, que em 1969 teria sido o primeiro a utilizar o termo, ainda consideram que em 1934, Otlet ao forjar o termo *bibliometrie* estava inclinado a conceber uma nova disciplina científica (ALVARADO, 2007, CHUEKE; AMATUCCI, 2015).

Quanto à fundamentação, a bibliometria é regida por três leis fundamentais, lei de Bradford, Zipf e Lotka, cada uma de acordo com sua medida abrange determinado aspecto acerca das produções científicas e possui sua própria teoria.

Bradford aferi quão reputável um periódico é, possibilitando detectar os periódicos mais importantes e que priorização a circulação de determinado tema em específico. Zipf quantifica a frequência de palavras-chave, revelando os temas mais recorrentes em uma linha de pesquisa. Lotka mensura a produtividade dos autores, identificando os autores mais produtivos em um campo do conhecimento (CHUEKE; AMATUCCI, 2015).

Para Bradford, também conhecida como lei da dispersão, aplica-se aos periódicos e tem como objetivo quantificar o número de publicações de um determinado tema em uma dada revista científica, assim podendo comparar com os demais canais de publicação e determinar quais periódicos são mais produtivos e relevantes em determinado campo do conhecimento (ALVARADO, 2007).

Ainda, os artigos estão distribuídos igualmente em três zonas de periódicos, muito produtivas, produtividade intermediária e pouco produtivas. Na zona produtiva há um pequeno número de periódicos que tem elevado rendimento e

tendem a exercer autoridade e influência em determinado campo de estudo (MACHADO JUNIOR et. al., 2016).

Na zona de eficiência intermediária encontra-se um número maior de revistas científicas menos produtivas e na terceira zona encontra-se muitos periódicos pouco produtivos.

Zipf ou lei do mínimo esforço quantifica a frequência de palavras-chave nas publicações, resultando uma lista ordenada de termos mais frequentes por assunto (RIBEIRO; MOLINA; OLIVEIRA, 2015). A teoria do mínimo esforço é composta de duas partes, na primeira parte as palavras que aparecem repetidamente representam o assunto do texto e a segunda parte, em que todas as palavras que menos se repetem tem a mesma frequência (MEDEIROS et. al., 2018, tradução nossa).

Lotka propõe que a maioria das publicações advêm da minoria dos autores, ou seja, estabelece uma relação inversa, conhecida como lei do quadrado inverso, em seus estudos Lotka identificou que a produção de poucos autores muito produtivos equiparava-se a toda a produção de muitos autores pouco produtivos (MACHADO JUNIOR et. al., 2016).

Portanto, estas são as leis basilares da bibliometria e as pesquisas são realizadas empregando uma lei de forma isolada ou um conjunto destas leis, conforme o objetivo da investigação.

A bibliometria e suas leis possibilitam a avaliação da produção científica e isto ocorre por meio de grupos de indicadores bibliométricos, como: qualidade científica, atividade científica, impacto científico e associações temáticas.

Qualidade científica é um índice obtido por meio da crítica realizada ao conteúdo de uma publicação, enquanto que atividade científica é uma medida, e portanto quantitativa acerca do, número de trabalhos publicados, número da produção individual dos autores, número de coautorias entre outros (LOPES Silvia et. al., 2012).

O indicador de impacto científico é composto dos seguintes subindicadores: impacto dos trabalhos e impacto das fontes.

O número de citações recebidas é um indicador dos trabalhos, já o fator de impacto das revistas, índice de citação imediata e a influência das revistas, são o que representam os indicadores de impacto das fontes.

Ainda, as associações temáticas são indicadas pelas análises de citações e análises de referências.

Além dos grupos supracitados, há os indicadores propriamente ditos.

O Fator de Impacto (FI) é um indicador de impacto dos periódicos, incluídos na *Science Citation Index* do *Institute* for *Scientific Information* (ISI) lidando com a frequência de citações de um artigo, realizando uma contabilização das citações efetuadas em um ano e então comparando com as citações de documentos publicados nos dois anos anteriores, o fator de impacto de uma revisa em 2019 seria dado conforme fórmula 6 e 7.

$$FI_n = \frac{citações_{(n-2)} + citações_{(n-1)}}{artigos_{(n-2)} + artigos_{(n-1)}} \quad (6)$$

$$FI_{2019} = \frac{citações_{2017} + citações_{2018}}{artigos_{2017} + artigos_{2018}} \quad (7)$$

Quanto à produtividade e impacto dos pesquisadores, é utilizado o indicador *h-index,* este índice trabalha com o número das citações dos artigos, lista-se todos os artigos de um autor ordenando-a decrescentemente por número de citações, quando existe um número máximo de citação que contempla a todas as citações então este é o *h-index* do autor. Se o *h-index* de um investigador é dez, indica que de todos os seus artigos o menos citado tem dez citações.

Outro indicador relativo aos periódicos é o e*igenfactor Metrics,* aplicado as revistas que utilizam os indicadores *Eigenfactor Score* (EF) e A*rticle Influence Score* (AI). O EF é baseado no número de citações do ano corrente dos artigos publicados nos cinco anos anteriores, avaliando a relevância da revista, enquanto que o AI determina quão influente um artigo é, na revista que o mesmo é publicado, durante os cinco anos posteriores a sua publicação, para determinar este valor o indicador *eigenfactor* da revista é dividido pelo total de artigos publicados da revista.

A pesquisa bibliométrica além das leis e indicadores conta com o auxílio de ferramentas para o desenvolvimento da investigação, as ferramentas de pesquisa bibliométrica em maior uso são *Web of Science, Scopus e Google Scholar Metrics,* por meio destas fontes o investigador científico tem a sua disposição para consulta indicadores como o *h-index*, asssim como inúmeros artigos de pesquisa. (LOPES Silvia et. al., 2012).

Finalmente, após contextualização e embasamento teórico relacionado a bibliometria, em que foi abordado os principais conceitos e fundamentos, assim como leis elementares deste campo de estudo e também ferramentas úteis ao desenvolvimento de uma pesquisa bibliométrica, o trabalho avança, e na sequência são descritos os passos metodológicos abordados para a realização da pesquisa bibliométrica acerca da produção de conhecimento em ciência da computação pelo emprego de análises implícitas.

5 TRABALHOS CORRELATOS

No processo de revisão bibliográfica não houve êxito em encontrar trabalhos muito semelhantes ao que propõe-se realizar, cujo propósito é realizar puramente um estudo bibliométrico onde as análises implícitas: árvores de decisão, classificadores *bayesianos*, redes neurais artificiais, máquinas de vetores de suporte, *k-means*, *k-medoid*, *DBSCAN*, Apriori e *FP-Growth*, são utilizadas na produção de conhecimento em ciência da computação.

Contudo foram encontrados artigos que utilizam a análise bibliométrica e que alguma das análises implícitas citadas compunham o trabalho, sendo assim estes foram relatados conforme segue.

5.1 AGRUPAMENTOS EPISTEMOLÓGICOS DE ARTIGOS PUBLICADOS SOBRE *BIG DATA ANALYTICS*

Furlan e Laurindo (2017) objetivando identificar os principais aspectos, correntes das publicações referente a *big data analytics* desenvolveram uma pesquisa bibliométrica na base de dados ISI *Web of Science*. Como estratégia de busca foi utilizada a seguinte expressão "*big data*" *OR* "*big data analytics*" *OR* "*big data analysis*", retornando 5.174 publicações, das quais foram mantidos 1.673 artigos, estes também foram filtrados por tema: manufatura, matemática, computação, negócios, economia e ciências sociais. Com auxílio do software *VosViewer* foi realizado levantamento das palavras mais utilizadas nos títulos dos artigos e representadas graficamente em *clusters* de palavras. Analisando o *cluster,* possíveis nichos são identificados, coleta e mineração de dados, análise de dados no ambiente computacional e temas que se referem ao futuro em torno de *big data.* Com base no número de citações de cada artigo, os quinze mais citados tiveram seus conteúdos analisados, destes o mais citado teve 68 citações. O índice de coautoria também foi explorado neste trabalho e constatou-se que os autores mais citados são os com menor quantia de pesquisas publicadas. A análise de conteúdo pode identificar cinco grupos de assuntos principais, evolução do *big data, g*estão,

negócios e estratégia, comportamento humano e aspectos socioculturais, mineração dos dados e geração de conhecimento e internet das coisas. A pesquisa foi realizada em 2015 e retornou em grande maioria publicações dos últimos quatro anos.

5.2 ANÁLISE DE SENTIMENTO E MINERAÇÃO DE OPINIÃO: UMA REVISÃO BIBLIOMÉTRICA DA LITERATURA

Ceci, Alvarez e Gonçalves, (2016), constatando a relevância da área da análise de sentimento e pretendendo compreender a sua relação com a área conhecida como mineração de opinião, realizaram um artigo científico, tendo como meta realizar uma revisão bibliométrica. As bases de pesquisa *Scopus* e *Web of Knowledge* foram utilizadas como fonte de dados, e nestas foram realizadas as pesquisas no período de julho de 2016 a setembro de 2016. Quanto a consulta, foi utilizado como estratégia de busca a seguinte sentença "*semantic analysis*" AND "*opinion mining*". A partir das buscas, cento e catorze artigos estavam disponíveis para download, ainda, estes estão distribuídos entre os anos de de 2006 a 2016. Quanto aos autores, foi identificado trezentos e vinte e cinco autores de trinta e oito países, sendo a China o país com maior número de autores, vinte e nove. Dos dez autores que mais publicaram, o menor número de publicações por autor foi três e o maior foi sete. Após fase de análise integral do conteúdo dos artigos pode-se quantificar por área de interesse que, dois artigos tinham foco em análise de emoção, sessenta e cinco artigos em análise de sentimento, sete artigos em classificação de sentimento, sete artigos em dicionário de sentimentos, trinta e um artigos em mineração de opinião e dois artigos em recuperação de opinião. Referente aos métodos mais utilizados na fase de classificação, em ordem de posição decrescente são: *Support Vector Machine* (SVM), *Naïve Bayes,* POS *Tagging*, Processamento de Linguagem Natural (PLN), *Latent Dirichlet Allocation* (LDA), *clusterização, Point Wise Mutual Information* (PMI), *Named Entity Recognition* (NER), *Fuzzy Logic, Random Forest*.

Ceci, Alvarez e Gonçalves, (2016), foram detectadas duzentas e setenta

e sete palavras-chave. A palavra-chave “*sentiment analysis*” se sobressai com oitenta e quatro citações e a palavra-chave “*opinion mining*”, aparece setenta e seis vezes. A pesquisa analisou o campo de estudo aos fundamentos de análise de sentimento e mineração de opinião, delineando um cenário com índices bibliométricos focados na detecção das publicações e propensão da literatura.

5.3 APLICAÇÃO DE REDES NEURAIS NO BRASIL: UM ESTUDO BIBLIOMÉTRICO

Alves et al. (2016), em virtude da notória evolução do uso das Redes Neurais artificiais (RNA) nas produções científicas em diversos campos do conhecimento, realizaram um estudo bibliométrico, com objetivo de verificar o volume de publicações nos últimos dez anos, quais as áreas de classificação das pesquisas e nos trabalhos relativos a finanças propõe-se aprofundar o estudo referente a objetivo, amostra, técnicas e resultados. As fontes de dados escolhidas, foram: Coordenação de aperfeiçoamento de Pessoal de Nível Superior (CAPES) e *Scientific Periodicals Electronic Library* (SPELL). Este trabalho teve as seguintes regras como filtros. Os artigos que compuseram a amostra foram filtrados por período, contemplando os últimos dez anos, de 2004 a 2013 e os artigos deveriam apresentar no seu título a expressão “redes neurais”. Uma vez aplicado o primeiro estágio da pesquisa, os artigos foram classificados nas seguintes áreas de aplicação: contabilidade e finanças; saúde e medicina; engenharia e manufatura; *marketing* e aplicações gerais.

Ainda, Alves et al. (2016), quanto aos resultados obtidos, oriundos do portal da CAPES e SPELL restaram cento e vinte e seis artigos. Relativo às áreas de atuação, os artigos distribuem-se quantitativamente, a área de aplicações gerais teve cinquenta e nove artigos, engenharia e manufatura vinte e quatro, contabilidade e finanças vinte e dois, saúde e medicina dezessete e marketing quatro artigos. Quanto a produção científica por ano, os autores observam aumento de 30% na produção científica entre a primeira metade do período e a segunda metade, evidenciando crescimento da produção, ainda, a média anual é de doze publicações. Ainda no último processo metodológico abordado na pesquisa os autores realizaram

uma síntese dos artigos que eram aplicados a contabilidade e finanças, sendo assim em vinte artigos foi apresentado de forma tabulada título, objetivo, autores e ano, ainda observaram que a maioria dos estudos tiveram melhores resultados com aplicação de RNA.

Assim, estes foram os trabalhos que mais convergem com nosso objetivo de pesquisa, na sequência será apresentada a metodologia abordada na pesquisa.

6 METODOLOGIA

Alcançar um fim desejado, requer que um caminho seja percorrido, este pensamento aplicado à investigação científica, implica que o resultado de uma pesquisa necessita um caminho pelo qual as tarefas de pesquisas trilharão, e este caminho é fornecido pela metodologia.

A metodologia possibilita a escolha dos processos de investigação dos fatos e a definição dos procedimentos utilizados para alcançar os objetivos do estudo. Auxiliando na identificação dos métodos adequados para coleta, formatação, análise, apresentação e arquivamento de dados (PRAÇA, 2015).

Este trabalho, é uma pesquisa bibliométrica direcionada a investigar a produção de conhecimento em ciência da computação pelo emprego das análises implícitas, logo a bibliometria é utilizada como referencial metodológico. Portanto trata-se de uma pesquisa quantitativa visto que tem como objetivo, medir, aferir e quantificar.

Para o desenvolvimento desta exploração científica foram definidas determinadas etapas de execução a fim de servir como fases norteadoras do processo metodológico desta pesquisa.

Etapas definidas:

a) escolha das análises implícitas;

b) área de concentração da aplicação das análises;

c) definição da estratégia de busca;

c) definição das bibliotecas eletrônicas;

d) pesquisa exploratória;

e) download e organização dos artigos;

f) análise superficial dos artigos;

g) análise detalhada dos artigos;

h) extração de dados;

i) análise e relacionamento dos dados levantados;

A primeira preocupação da pesquisa foi definir quais análises implícitas comporiam a pesquisa, e foi decidido que as seguintes técnicas seriam investigadas

(AMARAL, 2018): árvores de decisão, classificadores bayesianos, redes neurais artificiais, máquina de vetores de suporte, *k-means*, *k-medoid*, *DBSCAN*, Apriori e *FP-Growth*.

Atendido o primeiro item da lista, passasse a delimitar a pesquisa por área, então foi definido o campo de atuação, e em virtude do objetivo da pesquisa ser a produção do conhecimento em ciência da computação as buscas foram direcionadas para trabalhos que empregaram as análises em campos de estudos dentro da ciência da computação.

Uma pesquisa bibliométrica conta com uma estratégia de pesquisa que basicamente direciona as buscas servindo como um filtro pelo qual deseja-se obter determinados artigos, neste trabalho a estratégia é composta de duas partes, a análise empregada e o campo de atuação.

Com as análises e o campo já delimitados foram criadas as estratégias de busca. Havendo nove análises, logo foram criadas nove estratégias de busca, contudo foi necessário criar mais uma estratégia para contemplar o termo *“classifiers”,* da estratégia *"Bayesian Classification",* que em algumas bases de dados retornavam conteúdos ou com “classifiers” ou com “classification”. Cada estratégia ficou estruturada com o nome da análise em inglês entre aspas duplas, seguido do operador lógico *AND* mais o nome da área em inglês entre aspas duplas.

Portanto estas foram as estratégias empregadas, *"Decision trees" AND "Computer science", "Bayesian Classifiers" AND "Computer science", "Bayesian Classification" AND "Computer science", "Artificial neural networks" AND "Computer science", "Support Vector Machine" AND "Computer science", "K-means" AND "Computer science","K-medoid" AND "Computer science", "DBSCAN" AND "Computer science", "Apriori" AND "Computer Science"* e *"FP-Growth" AND "Computer Science".*

A próxima etapa é a escolha das bibliotecas digitais que foram pesquisadas, quanto a isto, por serem mais difundidas no meio científico, as três fontes de dados foram utilizadas: SciELo, Scopus e *Web of Science*.

A opção por estas bases de dados é devido serem as mais utilizadas em bibliometria, além disso são as que a UNESC tem convênio e também tem uma

vasta coleção de revistas indexadas, nos trabalhos correlatos é verificado que os autores utilizaram pelo menos uma destas bases, isto também corrobora com Lopes Silvia et al.(2012) que aponta Scopus e *Web of Science* como as ferramentas mais utilizadas em bibliometria.

As etapas anteriores, são pré-requisitos para realizar a pesquisa, logo estando concluídas, a pesquisa entra nas etapas práticas e é dado início à pesquisa propriamente dita.

Nas bases de dados foram aplicados os seguintes critérios de buscas, além da *string* de consulta, o único filtro aplicado é referente à artigos que sejam de acesso público, ou seja, sem custos ou restrições, também não foi restringido por data, autor ou qualquer outro limitador de busca.

Esta fase foi realizada na UNESC, no Laboratório de Pesquisa Aplicada em Computação e Métodos Quantitativos (LACOM), juntamente com membros deste laboratório já habituados a realizar pesquisas bibliométricas.

A pesquisa exploratória foi realizada nas datas compreendidas entre 08 e 11 de abril de 2019 e todas as estratégias de buscas foram executadas nas três bases de dados escolhidas. Os resultados de quantidades de artigos *X* estratégia *X* base de dados foram tabulados e serviram como base para a próxima etapa da pesquisa.

Todas as tabulações foram realizadas no *software online* Planilhas *Google.*

Os dados obtidos na etapa anterior foram utilizados para realizar o processo de *download* e organização dos arquivos, foram criadas três pastas com os respectivos nomes das bases de dados, Scielo, Scopus e *Web of Science* e dentro de cada diretório foi criada uma pasta para cada análise, com a estrutura de pastas montada, os arquivos foram organizados na pasta da base de origem do arquivo, na análise correspondente e colocado o número do arquivo como prefixo do nome, este número representa a quantidade de arquivos de cada análise em cada base.

Esta fase da pesquisa foi desenvolvida entre os dias 25 de abril de 2019 e 03 de maio de 2019.

Após fazer o *download* e organizar os arquivos foi desenvolvida uma tabela contendo as seguintes colunas, “periódico”, “análise”, “título do artigo”, “sim/não”, “aplicação” e “observação”. Então foi realizada a leitura superficial de cada artigo, que consiste da leitura do título, subtítulo, resumo, palavras-chave e título dos capítulos, feito isto os artigos foram registrados na tabela e obtido os seguintes dados, nomes dos periódicos, nomes dos artigos, o campo sim ou não refere-se ao enquadramento do artigo com nosso objetivo de pesquisa, o campo aplicação é referente a área de estudo que é identificada no artigo e o campo observação para qualquer informação adicional que fosse necessária para a pesquisa.

Quanto ao enquadramento do artigo, foi usado como critério estar evidente o uso de uma das análises na área da ciência da computação e o documento deve ser um artigo, livros ou artigos estritamente caracterizados como levantamento bibliográfico ou análises bibliométricas foram descartados para a próxima fase.

Ao fim da primeira triagem dos artigos, foi criada outra aba na tabela e copiado apenas os registros dos arquivos que tinham valor *sim*, na coluna sim/não e fossem aplicados à computação, então é iniciada a etapa de análise detalhada dos arquivos, que consiste da leitura completa dos artigos, nesta fase também foi avaliado o enquadramento do artigo.

Novamente, os registros que possuem o valor *sim* na coluna sim/não e fossem aplicados à computação foram mantidos para próxima fase, estes foram copiados para nova aba da tabela, que contém as seguintes colunas, “nº”, “autor”, “*h-index”*, “coautores”, “ano”, “universidade”, “país”, “revista”, “*qualis”*, “título”, “palavras-chave”, “objetivo”, “metodologia”, “resultados”, “limitação”, “conclusão”, “análise empregada” e “subcampo de aplicação”.

Nesta etapa algumas informações foram transcritas do artigo para a planilha, como, autor, coautores, ano, universidade, país, revista, título e palavras-chave, outras necessitaram análise e extração de informações como, objetivo, metodologia, resultados, limitação, conclusão e técnica empregada enquanto outras precisaram ser consultadas em fontes específicas como *h-index*, que é consultado

na base da Scopus, assim como o *qualis* que é consultado na plataforma Sucupira.

A última planilha representa os dados finais da pesquisa, a partir destes foi realizada a análise bibliométrica, aplicando as leis basilares da bibliometria, e índices bibliométricos pertinentes, possibilitando então, o entendimento das tendências das pesquisas da área assim como lacunas.

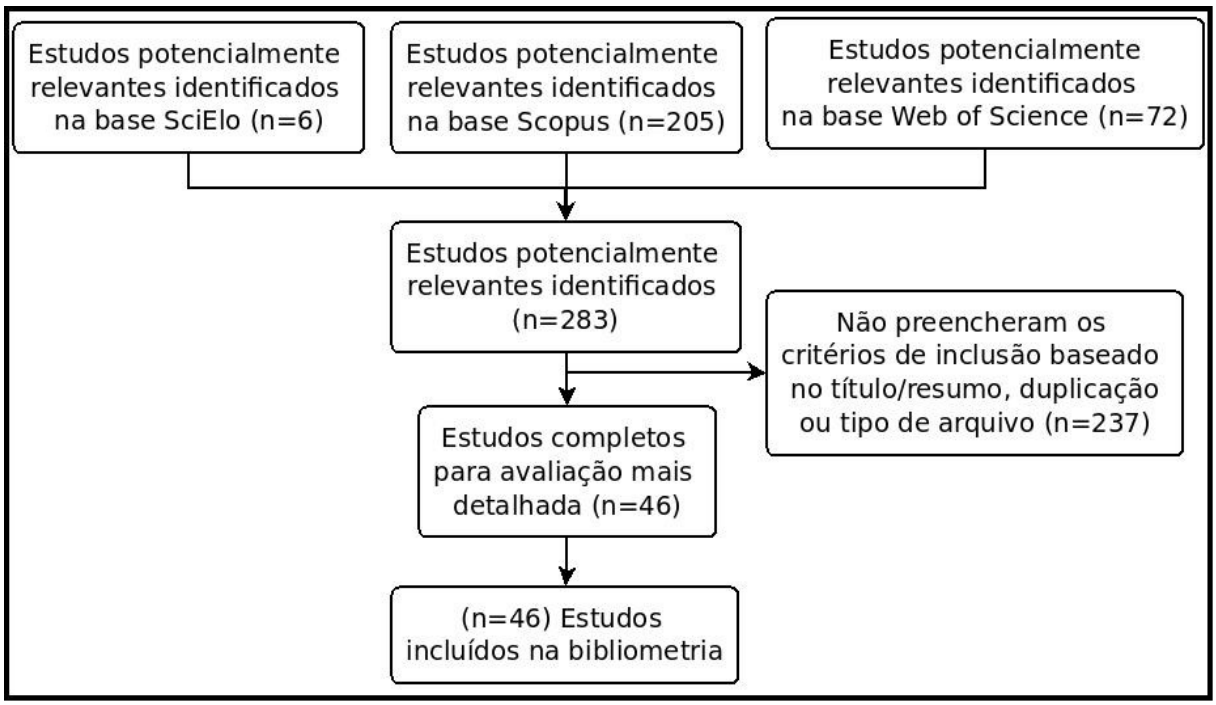
Quanto a visualização, apresentação e a discussão dos resultados, assim como das principais informações são tratados no capítulo seguinte.

7 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Neste capítulo é realizado a apresentação, análise e discussão dos resultados obtidos, procurando correlacioná-los com os trabalhos correlatos, assim como com a aplicação das leis basilares da bibliometria e os indicadores aplicáveis ao caso, ainda, é utilizado a tabulação de dados como método de apresentação.

De forma resumida a figura 12, contém informações dos artigos na primeira fase, e quantos restaram para as fases seguintes.

Figura 12 – Representação da quantia de arquivos em cada fase



Fontes: Dados da pesquisa (2019).

Conforme as etapas definidas na metodologia, após a estratégia de busca, campo de atuação e bases de dados estarem definidas, alguns dados já puderam ser obtidos nas fases iniciais da pesquisa. Os resultados primários retornados das *strings* de consulta podem ser visualizados na tabela 1.

Tabela 1 - Distribuição das análises implícitas nos repositórios *Scielo*, *Scopus* e *Web of Science*

Análise implícita	**Repositórios**			**Total**
	SciElo	*Scopus*	*Web of Science*	
Apriori	0	10	3	**13**
Artificial neural networks	3	63	20	**86**
Bayesian Classification	1	1	0	**2**
Bayesian Classifiers	0	2	1	**3**
DBSCAN	0	4	4	**8**
Decision Tree	2	34	9	**45**
FP-Growth	0	0	4	**4**
K-means	0	29	14	**43**
K-medoid	0	3	0	**3**
Support Vector Machine	0	59	17	**76**
10	**6**	**205**	**72**	**283**

Fontes: Dados da pesquisa (2019).

As estratégias de buscas retornaram os seguintes dados, são nove análises conforme tabela, no entanto durante a pesquisa observa-se que os métodos de classificação *bayesianos* eram indexados pelo termo *classifiers* e *classification,* por este motivo a pesquisa passou a ter dez estratégias de buscas para cobrir as nove análises investigadas.

Está fase resultou duzento e oitenta e três artigos divididos em análises e repositórios, observa-se que a maioria absoluta dos artigos estão indexados no repositório *Scopus,* com duzento e cinco artigos, aproximadamente 72,43% de todos os artigos registrados na primeira fase.

A superioridade númerica de artigos da *Scopus* confirmam o resultado obtido por Silva et al. (2013) em seu trabalho bibliométrico, no qual opta pela *Scopus* em face de ser mais númerosa em artigos, sendo que 74,1% dos artigos da pesquisa de Silva et al. (2013) estão indexados neste repositório.

Conforme consulta na página da *Scopus* disponível no endereço eletrônico https://www.Scopus.com/sources, em 02 de junho de 2019 é registrado 5393 periódicos abertos indexados nesta base.

Quanto as análises, destacam-se *Artificial Neural Network* com 86 artigos correspondendo a 30,38% de todos os artigos, seguido de *Support Vector Machine* com 76 artigos correspondendo a 26,85% de todos os artigos. Os artigos das duas análises somados representam a maioria absoluta de todos os artigos, por outro

lado, observa-se a análise com menor número de artigos, *k-medoid* com apenas três artigos, correspondendo a 1,06%.

Foi feito o download de todos os artigos que são representados pela tabela 1, e mediante leitura do título, resumo, palavras-chave e nome dos capítulos foram classificados, para a fase seguinte 46 artigos, 237 artigos foram excluídos da análise pelos motivos citados no capítulo de metodologia, apenas 19,40% dos artigos estavam alinhados com nossos objetivos de pesquisa e passaram para as próximas fases.

Tabela 2 - Distribuição das análises implícitas nos repositórios *Scielo*, *Scopus* e *Web of Science* após primeira fase

Análise implícita	**Repositórios**			**Total**
	SciElo	*Scopus*	*Web of Science*	
Apriori	0	1	1	**2**
Artificial neural networks	1	12	0	**13**
Bayesian Classification	0	0	0	**0**
Bayesian Classifiers	0	0	0	**0**
DBSCAN	0	1	0	**1**
Decision Tree	0	7	0	**7**
FP-Growth	0	0	1	**1**
K-means	0	8	1	**9**
K-medoid	0	1	0	**1**
Support Vector Machine	0	12	0	**12**
10	**1**	**42**	**3**	**46**

Fontes: Dados da pesquisa (2019).

Na tabela 2, consta a distribuição dos artigos, que corroboram com a tabela 1 e acentuam a quantidade de artigos do repositório *Scopus*, que mantém quarenta e dois dos quarenta e seis artigos remanescentes, representando 91,30% de todos os artigos.

Referente aos artigos remanescentes, quarenta e seis, o título e link de acesso destes encontram-se no apêndice a.

Quanto as análises, destacam-se *Artificial Neural Networks* e *Support Vector Machine* com treze e doze artigos respectivamente, mantendo juntas o predomínio dos artigos. Também observa-se que nenhum artigo de classificadores *bayesianos* está presente na fase final da pesquisa.

Na última fase da pesquisa, temos quarenta e seis artigos e quarenta e

cinco autores, apenas um autor tem dois artigos presentes nesta fase, o pesquisador Chinês Ye, Yongkai, filiado a *National University of Defense Technology* , seus dois trabalhos empregam a técnica de análise implícita *k-means.* Este autor tem *h-index* dois, segundo consulta a base *Scopus.*

Referente aos autores, quanto ao *h-index,* dos quarenta e cinco autores, dois destes não estão indexados na *Scopus h-index,* Andrade, Guilherme e Nibir, Nayan Bora, para os demais autores, é representado na tabela 3.

Tabela 3 – Relação de *h-index* dos autores

Intervalo *h-index*	**Frequência**
0 |-- 2	11
2 |-- 4	11
4 |-- 6	9
6 |-- 8	4
8 |-- 10	5
10 |-- 12	1
12 |-- 14	2
14 |--16	1
Autores	**43**

Fontes: Dados da pesquisa (2019).

A tabela 3 representa os *h-index* distribuídos em classes, com amplitude de duas unidades de índice em cada classe, e a frequência é a quantia de autores em cada classe, os dados nos mostram que poucos autores têm *h-index,* elevado e que muitos autores têm baixo *h-index,* os três autores com alto *h-index,* são Brázdil Thomáš, com *h-index* quinze, Artur S. D'Avila Garcez com *h-index* treze e Mahajan, Meena com *h-index* doze, estes três autores são dos respectivos Países, República Tcheca, Inglaterra e Índia.

O *h-index* é um indicador de impacto e produtividade do autor, e o resultado obtido conforme tabela 3 confirma a lei de Lotka, que propõe que a maioria das publicações advêm da minoria dos autores (MACHADO JUNIOR et. al., 2016).

Ainda, analisando o cenário de autoria e coautoria, os quarenta e seis artigos representam o trabalho de cento e cinquenta e quatro pesquisadores, sendo quarenta e cinco autores e cento e nove coautores, em média cada obra envolve três pesquisadores, um autor, e dois coautores. Entre os coautores, apenas Xinwang Liu, Qiang Liu e Jianping Yin participam em mais de um trabalho, analisando este

dado em específico, foi identificado que estes três autores publicam juntamente com o autor Ye, Yongkai, e, estes trabalhos têm como foco a análise *k-means,* tanto autores quanto coautores são Chineses.

Esta informação, de autores publicando juntos em um determinado tema, neste caso agrupamentos, *clusterização* e a técnica de *k-means,* nos motivou a realizar breve pesquisa nas demais publicações de Ye, Yongkai, e constata-se que, nas demais pesquisas, a técnica *k-means* também é abordada, e o autor mantém a relação de parceria com os coautores dos trabalhos que fazem parte desta pesquisa.

As informações levantadas estão disponíveis na página da *Scopus* no seguinte endereço eletrônico https://www.Scopus.com/authid/detail.uri?authorId=57194193216. A bibliometria também pode ser observada e medida pela ótica social, quando autores, instituições se unem para investigações científicas na forma de redes de cooperação (SEHNEM et. al., 2013).

Além disso, analisando os artigos quanto aos coautores, identifica-se que a grande maioria dos trabalhos têm um ou dois coautores e apenas uma pequena parte têm a rede de coautoria mais ampla com quatro ou cinco pesquisadores, a tabela 4 apresenta estes relacionamentos.

Tabela 4 – Número de coautores por artigo

Nº coautores	**Nº artigos**
1	14
2	14
3	9
4	5
5	4
Total	**46**

Fontes: Dados da pesquisa (2019).

Quanto a produtividade por ano, conforme tabela 5 o ano de 2018 tem a maior quantidade de artigos, dezesseis, nos demais anos a produtividade não é tão acentuada quanto a dos anos de 2013 e 2017 têm seis artigos em cada ano, já em 2012 tem quatro artigos. Observa-se que a produtividade anual anterior a 2018 tem leves oscilações no entanto o ano de 2018 apresenta notável crescimento.

Tabela 5 – Produtividade por ano.

Ano	Número artigos
2019	2
2018	16
2017	6
2016	3
2015	1
2014	1
2013	6
2012	4
2011	1
2009	2
2007	1
2006	1
2005	2
Total	**46**

Fontes: Dados da pesquisa (2019).

Foram analisados 46 artigos de 42 universidades, logo entende-se que poucas universidades têm mais de um artigo na lista, estas instituições são, *Hassan 1st University Settat* do Marrocos com dois artigos, *National University of Defense Technology* da China com dois artigos, *Renmin University of China,* também da China com dois artigos e *University Rabat* do Marrocos também com dois artigos, o restante das universidades teve apenas um artigo cada.

Analisando a produtividade por país, destacam-se países como a China e a Índia, que mais contribuíram com a pesquisa, a tabela 6 agrupa os países pela produtividade, o grupo com número de artigo um incluí os países com um artigo nesta pesquisa, desta forma consegue-se observar que a produção da China e da Índia se equipara a de todos os países do grupo um, novamente o princípio de Lotka é observado, poucos produzindo muito e muitos com pouca produção.

Tabela 6 – Produtividade por País

Nº artigos	Países
9	China
7	Índia
4	Marrocos
2	Alemanha, Estados Unidos da América, República da Coreia (Coreia do Sul), Romênia e Vietnã
1	Arábia Saudita, Brasil, Colombia, Dinamarca, Filipinas, Grécia, Indonésia, Inglaterra, Irã, Malásia, México, Palestina, República Tcheca, Rússia, Sri Lanka e Turquia

Fontes: Dados da pesquisa (2019).

Quanto aos periódicos, os mesmos são avaliados quantitativamente e qualitativamente, quantitativamente destacam-se os periódicos *Hindawi, International Journal of Advanced Computer Science and Applications* e *IEE access* com oito, cinco e quatro artigos respectivamente conforme tabela 7.

Qualitativamente a tabela 8, traz a relação de *qualis* dos artigos, observa-se que muitas revistas não possuem avaliação *qualis* na plataforma Sucupira.

Tabela 7 – Relação artigos X Periódicos

Nº artigos	**Nome periódico**	**Nº periódico**
8	Hindawi	1
5	International Journal of Advanced Computer Science and Applications	1
4	IEEE access	1
3	International Journal of Machine Learning and Computing; Procedia Computer Science; Theoretical Computer Science	3
2	Information Technology Journal	1
1	Advances in Electrical and Computer Engineering; De Gruyter Open; Heliyon; IBM India Research Lab; Ingeniería y Universidad; International Conference on Computational Science; International Conference on Image Analysis and Processing; International Federation for Information Processing; Journal of Computing and Information Technology; Journal of Information Processing Systems; Journal of Mathematical Neuroscience; Lecture notes in computer science (internet); Materials Science and Engineering; PHYSICAL REVIEW X; Procedia Engineering; Telkomnika; Transactions on Architecture and Code Optimization; Turkish Journal of Electrical Engineering & Computer Sciences	18

Fontes: Dados da pesquisa (2019).

Tabela 8 – Relação *Qualis* X periódico X artigos

Qualis	Periódicos	Nº periódicos
A1	Theoretical Computer Science	1
B1	Advances in Electrical and Computer Engineering; Transactions on Architecture and Code Optimization	2
B3	IEEE access	1
C	Procedia Computer Science; Lecture notes in computer science (internet)	2
[S.l]	Hindawi; International Journal of Advanced Computer Science and Applications; International Journal of Machine Learning and Computing; Information Technology Journal; De Gruyter Open; IBM India Research Lab; Ingeniería y Universidad; International Conference on Computational Science; International Conference on Image Analysis and Processing; International Federation for Information Processing; Journal of Computing and Information Technology; Journal of Information Processing Systems; Journal of Mathematical Neuroscience; Materials Science and Engineering; PHYSICAL REVIEW X; Procedia Engineering; Telkomnika; Turkish Journal of Electrical Engineering & Computer Sciences; International Journal of Machine Learning and Computing	20

Fontes: Dados da pesquisa (2019).

Estabelecendo uma relação entre revistas e artigos que compõem a pesquisa, é possível confirmar alguma proximidade com a lei de Bradford, que propõe zonas de produtividade. Nesta pesquisa a zona muito produtiva é composta por duas revistas com cinco e oito artigos cada, a zona com produtividade razoável é formada por quatro revistas que possuem quatro ou três artigos e finalmente a zona menos produtiva com dezenove periódicos com um ou dois artigos, estabelecendo assim uma relação com que Bradford afirma (ALVARADO, 2007).

A relação *qualis* dos artigos é verificada na tabela 9, observa-se que apenas três artigos foram públicados em revistas com conceito máximo, ou seja A1, e que a maioria absoluta dos artigos não tem índice *qualis*, este número representa 71,73% dos artigos.

Tabela 9 – Relação *qualis* X artigos

Qualis	**Nº artigos**
A1	3
B1	2
B3	4
C	4
[S.I]	33
Total	**46**

Fontes: Dados da pesquisa (2019).

Os artigos também foram analisados quanto as palavras-chave, nesta pesquisa foram catalogadas as palavras-chave que estão explicitamente declaradas, em treze artigos não foram encontradas as palavras-chave.

Das palavras catalogadas destaca-se *clustering, k-means, support vector machine, computing, network, mining, system, tree, artificial neural network e classification* que são representadas na tabela 10.

O fundamento bibliométrico que trata das palavras chaves é a lei de Zipf, que entende que as palavras mais frequentes tendem a representar o assunto do texto (RIBEIRO; MOLINA; OLIVEIRA, 2015). Não foi aplicado este pressuposto a um texto específico, mas, as palavras-chave mais frequentes de todos os artigos foram relacionadas com o que foi classificado como grupo de pesquisa, que está mencionado nos capítulos seguintes.

E estas palavras com maior frequência correspondem com os temas mais pesquisados assim como com as análises implícitas mais empregadas.

Tabela 10 - palavras-chave

Frequência	Palavra-chave
8	Clustering
7	K-means
7	Support vector machine
6	Computing
6	Network
5	Mining
5	System
5	Tree
4	Artificial neural network
4	Classification

Fontes: Dados da pesquisa (2019).

Conforme a análise dos artigos, pode-se identificar grupos de artigos com objetivos semelhantes, foi identificado cinco grupos nos quais os artigos se enquadram.

Grupo de artigos que de forma geral o objetivo da pesquisa esta relacionada a melhoria de algum problema de performance de algoritmo, ou pesquisa de aperfeiçoamento de desempenho em arquitetura computacional ou a correção de algum processo de algoritmo ou arquitetural da computação, é entendido como: pesquisa e desenvolvimento. Estas características estão presentes em vinte e seis artigos.

Quanto a análise adotada por este este grupo, são abordadas pelo menos uma destas técnicas, RNA, árvores de decisão, *k-means, SVM, DBSCAN,* Apriori e Fp-growth.

Há também trabalhos voltados para o Processamento de Linguagem Natural (PLN), estes trabalhos tem como foco desenvolver tecnologias que permitam que a interação homem-máquina seja mais transparente e natural, possibilitando às máquinas reconhecer as mais diversas formas pelas quais os seres humanos, se comunicam e se expressão, doze artigos compõem este grupo e as análises utilizadas são árvores de decisão, RNA e SVM.

A segurança computacional também é abordada por alguns artigos, cinco artigos estão alinhados com este objetivo de pesquisa e utilizam árvores de decisão, RNA e Apriori.

Detectado que um artigo tem como objetivo contornar problemas

relacionados as lacunas de dados em *datasets*, considerado um problema para as análises de dados e utiliza árvores de decisão para propor soluções neste campo.

Finalmente o quinto grupo está relacionado à pesquisa e indexação de conteúdo, dois artigos têm suas pesquisas voltadas a este propósito e utilizam *DBSCAN* e *k-medoid* nas soluções propostas.

Estas informações estão relacionadas na tabela 11, onde é associado grupos de pesquisa com quantia de artigos e as análises implícitas utilizadas.

Tabela 11 – Grupos de pesquisa relacionados a artigos e a análises implícitas.

Grupo de pesquisa	**Nº artigos**	**Análises implícitas**
Pesquisa e desenvolvimento	26	RNA, árvores de decisão, *k-means, SVM, DBSCAN,* Apriori e Fp-growth.
Processamento de Linguagem Natural	12	Árvores de decisão, RNA e Apriori.
Segurança computacional	5	Árvores de decisão, RNA e Apriori.
Pesquisa e indexação de conteúdo	2	*DBSCAN* e *k-medoid.*
Ausência de dados	1	Árvores de decisão

Fontes: Dados da pesquisa (2019).

Conforme tabela 11, o grupo com maior número de artigos esta relacionado a melhorias, evoluções e correções de tecnologias computacionais por meio de alguma análise implícita, são vinte e seis artigos de um total de quarenta e seis, ou seja 56,52% das pesquisas concentram esforços, em pesquisa e desenvolvimento computacional.

Enquanto que doze artigos estão relacionados a PLN, ou seja 26% do total de artigos, segurança computacional é representada por cinco artigos que é equivalente a 10,8%, ausência de dados em *datasets* tem apenas um artigo, sendo igual 2,17% e finalmente pesquisa e indexação de conteúdo com dois artigos representa 4,34% dos artigos.

Quanto as análises implícitas, pela perspectiva de grupos de pesquisa, as árvores decisão são empregadas em quatro dos cinco grupos, sendo a análise mais utilizada, Apriori e RNA são empregadas em três grupos, *DBSCAN* é utilizada em

dois grupos e é utilizada em apenas um grupo as análises *k-means, k-medoid, SVM e FP-Growth.*

Finalmente, o último indicador bibliométrico analisado é quanto ao idioma das pesquisas, constatou-se que todos os artigos desta pesquisa estão na língua inglesa.

8 CONCLUSÃO

Esta pesquisa teve por objetivo mapear trabalhos com produção de conhecimento voltada para ciência da computação por meio de alguma destas análises implícitas, árvores de decisão, classificadores *bayesianos,* redes neurais artificiais, máquina de vetores de suporte, *k-means, k-medoid, DBSCAN*, Apriori e *FP-Growth.*

Para o melhor entendimento e desenvolvimento desta investigação científica também foi realizado levantamento bibliográfico sobre ciência dos dados, análises implícitas e bibliometria, a revisão bibliográfica destes assuntos foi apresentada em capítulos específicos.

Ciência dos dados foi desenvolvido no capítulo dois, em que é resgatado sua história e também é tratado sobre o dado e seu ciclo de vida, no capítulo posterior é dissertado sobre o conceito de análises implícitas, e também estas análises são fundamentadas, de forma que em um único capítulo o conceito de análise implícita esta relacionado com as técnicas algorítmicas de análise de dados, que até então na literatura encontra-se associada a data *mining,* descoberta de conhecimento, *machine learning* e outros termos.

A ferramenta metodológica que possibilitou o desenvolvimento deste trabalho, também foi abordada em um capítulo exclusivo, em que é relatado suas origens e seus fundamentos a fim de nos capacitar ao entendimento e condução de um estudo bibliométrico.

Como método de pesquisa foi utilizado a bibliometria, resultando em diversos indicadores qualitativos e quantitativos, conforme estava proposto no trabalho.

Visto que todos os artigos analisados na última fase da pesquisa estão no idioma inglês, observa-se que na pesquisa, inglês é o idioma dominante e entende-se que a fluência neste idioma para um pesquisador é algo desejável.

Quando foi realizada as pesquisas em busca dos artigos, não foi aplicado filtro de data a fim de aumentar a abrangência da linha do tempo das publicações e então identificar quão evoluída estão as pesquisas.

Na nossa coleção de artigos remanescentes, cronologicamente os dois primeiros artigos são do ano de 2005 e os dois últimos do ano de 2019, logo a pesquisa representa quinze anos de produção científica, a média de produção anual é de aproximadamente três artigos por ano.

Nos anos de 2005 a 2011 foram produzidos sete artigos, em média um por ano, de 2012 a 2013 teve produtividade de dez artigos, sendo em média cinco por ano, observa-se que até este ponto a produtividade de artigos esta crescendo, nos anos de 2014 a 2017 foram produzidos onze artigos, sendo em média 2,75 artigos por ano, no entanto nos anos de 2018 a 2019 foram produzidos dezoito artigos com média de nove artigos por ano, tendo nesta época um pico de crescimento, no entanto dezesseis artigos são do ano de 2018 e apenas dois artigos são do ano de 2019, vale lembrar que os artigos do ano de 2019 são apenas dos primeiros meses.

Baseado nos dados de produção anual, entende-se que de 2005 até 2018 esta linha de pesquisa está em crescimento.

Referente aos autores, é destacado o pesquisador Chinês Ye Yongkai por ser o único com dois artigos nesta pesquisa, todos os demais autores tem apenas um artigo, além disso Ye Yongkai publica juntamente com os pesquisadores Xinwang Liu, Qiang Liu e Jianping Yin e analisando outros artigos deste autor foi identificado que além de manter a relação com seus coautores suas pesquisas são endereçadas a problemas que envolvem agrupamentos, *clusterização* e emprego de *k-means.*

Em virtude destas evidências, o mesmo é destacado entre os autores, contudo seu *h-index* ser dois, considerado baixo em relação aos três autores com os maiores *h-index* desta pesquisa, sendo Brázdil Thomáš, com *h-index* quinze, Artur S. D'Avila Garcez com *h-index* treze e Mahajan, Meena com *h-index* doze, estes três autores são dos respectivos países, República Tcheca, Inglaterra e Índia.

Quanto a produtividade por país, destaca-se China e Índia como sendo os países com maior número de produções, nove e sete respectivamente.

Os temas abordados nos artigos desta pesquisa são representados pelas seguintes palavras-chave, *clustering*, *k-means*, *support vector machine*, *computing*,

network, *mining*, *system*, *tree*, *artificial neural network* e *classification.*

As análises mais utilizadas são, árvores de decisão, redes neurais artificiais e *Apriori,* também foi observado que inexiste um artigo com classificadores *bayesianos* na fase final da pesquisa.

Também observa-se que os artigos podem ser classificados em cinco grupos de linhas de pesquisa, destas, duas possíveis tendências são destacadas: pesquisa e desenvolvimento computacional e processamento de linguagem natural. Ainda, uma provável lacuna, ausência de dados em *datasets*.

Referente aos periódicos, 71,73% dos artigos foram publicados em revistas sem avaliação *qualis,* ainda, a revista com maior número de artigos foi a Hindawi, que não tem avaliação *qualis,* por outro lado a revista com avaliação máxima, A1 é a *Theoretical Computer Science,* com a apenas um artigo na pesquisa.

Baseado nos dados da pesquisa os parágrafos anteriores denotam nossa conclusão e entendimento dos resultados obtidos. Para futuras pesquisas, a partir destas informações, outras investigações com direcionamento mais específico podem ser realizadas, por exemplo, é identificado uma possível carência de pesquisa referente ao uso de análises implícitas no contorno das lacunas de dados em datasets, por outro lado, também possibilitam explorações focadas nas duas áreas que mais concentram estudos, pesquisa e desenvolvimento computacional utilizando análises implícitas e processamento de linguagem de natural.

Também entende-se que em trabalhos futuros outros indicadores bibliométricos podem ser explorados, como fator de impacto das revistas. Ainda, fazer uso de tecnologias como *text mining*, para minerar os textos dos artigos.

Uma limitação deste trabalho ocorre devido ao fato de não terem sido encontrados outros trabalhos com o mesmo objetivo deste, assim não há uma referência comparativa, tanto para metodologia quanto para os resultados obtidos.

Também entende-se que não foi explorado os recursos dos softwares bibliométricos como *VosViewer, EndNote* e outros, que poderiam ter contribuído com análises mais profundas e geração de melhores resultados em termos de gráficos.

REFERÊNCIAS

ADAZIMA, K. R; BUSTAMAN, A; ALDILA, A. The implementation of k-means partitioning algorithm in HOPACH, clustering method. **Earth and Environmental Science.** Vol 243, apr/2019. Disponível em: https://iopscience.iop.org/article/10.1088/1755-1315/243/1/012073/meta . Acesso em: 24 mai. 2019.

AGARWAL, Ritu; DHAR, Vasant. Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. **Information Systems Research**, [s.l.], v. 25, n. 3, p.443-448, set. 2014. Disponível em: https://pubsonline.informs.org/doi/full/10.1287/isre.2014.0546. Acessoo em: 20 Abr. 2019.

AHMED, Faisal et al. **Classification of crops and weeds from digital images: A support vector machine approach**. Crop Protection, v40, p98-104, 2012. Disponível em: https://www.sciencedirect.com/science/article/pii/S026121941200124X . Acessoo em: 10 mai. 2019.

ALVARADO, R. U. A Bibliometria: história, legitimação e estrutura. **Para entender a ciência da informação**. Salvador: EDUFBA, 2007. Disponível em: https://www.academia.edu/1390400/A_BIBLIOMETRIA_HISTORIA_LEGITIMA%C3%87%C3%83O_E_ESTRUTURA . Acesso em: 05 mai. 2019.

ALVES et al. **Aplicação de redes neurais no Brasil: um estudo bibliométrico**. Biblionline n 12, p101-112, 2016. Disponível em: http://www.periodicos.ufpb.br/index.php/biblio/article/view/27738 . Acesso em 10 mai. 2019.

AMARAL, Fernando. **Introdução à ciência de dados: mineração de dados e big data**. Rio de Janeiro: Alta Books, 2016.

BINOTI, Daniel; LEITE, Helio; BINOTI, Mayara Luiza Marques da Silva. Configuração de Redes Neurais Artificiais para estimação do volume de árvores. **Ciência da Made**
ira (Braz. J. Wood Sci). Pelotas, v.05, n. 01, p. 58-67, Maio 2014 ISSN: 2177 -6830. Disponível em: https://periodicos.ufpel.edu.br/ojs2/index.php/cienciadamadeira/index . Acesso em: 17 mar. 2019.

BOGONI, Mariella A; MENEZES, Daiana G; FREITAS, Marina S. D. Estudo e aplicação de Regressão Logística usando R Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 6, n. 2, 2018. Disponível em: https://proceedings.sbmac.org.br/sbmac/article/view/2592. Acessoo em: 16 mar. 2019.

BONIDIA, Robson; BRANCHER, Jacques; BUSTO, Rosangela. Data Mining in Sports: A Systematic Review. **IEEE Latin America Transactions**. [S.I.]v 16. p232-239. 10.1109/TLA.2018.8291478. 2018.. Disponível em: https://www.researchgate.net/publication323198458_Data_Mining_in_Sports_A_Systematic_Review/download. Acesso em 20 mar. 2019.

BOTELHO, Fernando Rigo; RAZZONI Edelvino Filho. Conceituando o termo Business Intelligence: Origem e Principais Objetivos. **Sistemas, Cibernética e Informática**. [SI]v. 11, n. 1, 2014. Disponível em: http://www.iiisci.org/journal/CV$/risci/pdfs/CB793JN14.pdf. Acesso em: 13 maio. 2019.

BRITO, José André de Moura; SEMAAN, Gustavo da Silva; BRITO, Luciana Roque. Resolução do problema dos k-medoids via algoritmo genético de chaves aleatórias viciadas. **Anais do Xvii Simpósio de Pesquisa Operacional e Logística da Marinha**, [s.l.], p.50-61, ago. 2014. Editora Edgard Blücher. Disponível em: https://www.researchgate.netpublication269203283_RESOLUCAO_DO_PROBLEMA_DOS_KMEDOIDS_VIA_ALGORITMO_GENETICO_DE_CHAVES_ALEATORIAS_VICIADAS . Acesso em: 05 mai. 2019.

BUFREM, Leilah Santiago et al. Produção Internacional Sobre Ciência Orientada a Dados: análise dos termos Data Science e E-Science na Scopus e na Web of Science. **Informação & Informação**, Londrina, v. 21, n. 2, p.40-67, dez. 2016. Universidade Estadual de Londrina. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/26543/20114 . Acesso em : 21 Abr. 2019

CASSIANO, Keila Mara; PESSANHA, José F, Moreira. Análise espectral singular com clusterização baseada em densidade na modelagem de séries temporais. **XLVI Simpósio Brasileiro de pesquisa operacional. Pesquisa Operacional na Gestão da Segurança Pública** . [S.I.] set/2014 p 1287-1298. 2014. Disponível em: http://www.senaicimatec.com.br/wp-content/uploads/2017/03/renataesquiveldissertacao15junh12.pdf . Acesso em: 15 mai. 2019.

CASTRO, Ricardo Ferreira Vieira. **Análise de desempenho dos algoritmos Apriori e Fuzzy Apriori na extração de regras de associação aplicados a um Sistema de Detecção de Intrusos**. 2014 . f. 137. Dissertação (Mestre em Engenharia Eletrônica) - Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2014. Disponível em: http://www.pel.uerj.br/bancodissertacoes/Dissertacao_Ricardo_Ferreira_Vieira_de_Castro.pdf . Acesso em: 20 mai. 2019.

CASTRO, Rui Pedro Gomes. **“Canibalização” de produtos - um estudo**. 2016. f.100. Tese (Mestre em Engenharia de Sistemas) - Universidade do Minho, Braga - PT, 2016. Disponível em: http://hdl.handle.net/1822/41841 . Acesso em: 18 mai. 2019.

CECI, Flávio; ALVAREZ, Guilherme Martins, GONÇALVES, Alexandre Leopoldo. Análise de sentimento e Mineração de Opinião: uma revisão bibliométrica da literatura. **ESPACIOS**. v.38, n.14, p.12 out/2016. Disponível em: https://www.revistaespacios.com/a17v38n14/a17v38n14p12.pdf . Acesso em: 30 abr. 2019.

CHIAVEGATTO FILHO, Alexandre Dias Porto. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. **Epidemiol. Serv. Saúde**, Brasília , v. 24, n. 2, p. 325-332, June 2015 . Disponível em http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2237-96222015000200325&lng=en&nrm=iso>. accesso em 04 abr. 2019.

CHUEKE, Gabriel Vouga; AMATUCCI, Amatucci. O que é bibliometria? Uma introdução ao Fórum. **Revista Eletrônica de Negócios Internacionais**, São Paulo, v. 10, n. 2, p. 1-5, mai./ago. 2015. Disponível em: http://internext.espm.br/internext/article/view/330 . Acesso em: 21 mar. 2019.

COSTA, Evandro et al. Mineração de dados Educacionais: Conceitos, Técnicas, Ferramentas e aplicações. **Jornada de Atualização em Informática na Educação**. 2012. Disponível em: http://www.br-ie.org/pub/index.php/pie/article/view/2341 . Acesso em: 22 mai. 2019.

CUNHA, Lorena F. F. **Classificação de empresas em grupos de retorno das ações utilizando redes neurais artificiais**. 2018. f 42. Trabalho de Conclusão de Curso (Graduação em Gestão da Informação) - Universidade Federal de Uberlândia, Uberlândia, 2018. Disponível em: https://repositorio.ufu.br/handle/123456789/23226 . Acesso em:13 mai. 2019.

CURTY, Renata Gonçalves; CERVANTES, Brígida Maria Nogueira. Data Science: Ciência orientada a dados. **Informação & Informação**, Londrina, v. 21, n. 2, p.1-4, dez. 2016. Universidade Estadual de Londrina. Disponível em: www.uel.br/revistas/uel/index.php/informacao/article/download/27929/20119 . Acesso em: 02 mai. 2019.

DAI, Wei; JI Wei. A MapReduce Implementation of C4.5 Decision Tree Algorithm. **International Journal of Database Theory and Application.** Vol.7, Nº.1, pp.49-60. 2014. Disponível em: https://www.researchgate.net/publication/284467402_A_MapReduce_Implementation_of_C45_Decision_Tree_Algorithm/citation/download . Acesso em: 23 mai. 2019.

DIZON, Franz Stewart V et al. Learning of High Dengue Incidence with Clustering and FP-Growth Algorithm using WHO Historical Data. **Computing Research Repository (CoRR).** v. 1901. n. 11376. 2019. Disponível em: https://arxiv.org/abs/1901.11376 . Acesso em: 26mai. 2019.

DONOHO, David. 50 Years of Data Science. **Journal Of Computational And Graphical Statistics**, [s.l.], v. 26, n. 4, p.745-766, 2 out. 2017. Disponível em: https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf . Acesso em: 04 abr. 2019.

DOSCIATTI, Mariza; PATERNO, Lohann; PARAISO, Emerson. Identificando Emoções em Textos em Português do Brasil usando Máquina de Vetores de Suporte em Solução Multiclasse. 2013. Disponível em: https://www.researchgate.net/publication/277813389_Identificando_Emocoes_em_Textos_em_Portugues_do_Brasil_usando_Maquina_de_Vetores_de_Suporte_em_Solucao_Multiclasse . Acesso em: 15 mar. 2019.

El ARASS, Mohammed; TIKITO, Iman; SOUISSI, Nissrine. **Data lifecycles analysis: Towards intelligent cycle**. Abr 2017. Disponível em: https://www.researchgate.netpublication316191501_Data_lifecycles_analysis_Towards_intelligent_cycle . Acesso em: 27 mai. 2019.

ELGAML, Elsayeda et al. Improved FP-Growth Algorithm with Multiple Minimum Supports Using Maximum Constraints. **World Academy of Science, Engineering and Technology, International Science Index 101**. Dubai 2015. Disponível em: https://www.researchgate.net/publication/280611828_Improved_FP-Growth_Algorithm_with_Multiple_Minimum_Supports_Using_Maximum_Constraints/download . Acesso em: 20 abr. 2019.

EMC, Education Services; **Data Science e Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.** 2015.

FAGUNDES, Priscila Basto; MACEDO, Douglas Dyllon Jeronimo de; FREUND, Gislaine Parra. A produção científica sobre qualidade de dados em big data: um estudo na base de dados Web of Science. **Revista Digital de Biblioteconomia e Ciência da Informação**, [s.l.], v. 16, n. 1, p.194-210, 9 nov. 2017. Universidade Estadual de Campinas. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8650412 . Acesso em: 22 abr. 2019.

FAJAR, Muhammad; RACHMAD, Sri Hartini. Inflation Forecasting By Hybrid Singular Spectrum Analysis-Multilayer Perceptrons Neural Network Method. **10th ECB Workshop on "Forecasting Techniques: Economic Forecasting with Large Datasets"**. Frankfurt Germany 2018. Disponível em: https://www.researchgate.net/publication/326964762_Inflation_Forecasting_By_Hybrid_Singular_Spectrum_Analysis-Multilayer_Perceptrons_Neural_Network_Method/download . Acesso em: 27 abr. 2019.

FÁVERO, Patrícia Belfiore; ZOUCAS, Augusto Mollik. Redes neurais para previsão da produção industrial de diferentes segmentos. **Produto e Produção**, vol 17, n2, p-53-70, 2016. Disponível em: https://seer.ufrgs.br/ProdutoProducao/article/view/51900 . Acesso em: 10 mai. 2019.

FURLAN, Patricia Kuzmenko; LAURINDO, Fernando José Barbin. Agrupamentos epistemológicos de artigos publicados sobre big data analytics. **Transinformação**, [s.l.], v. 29, n. 1, p.91-100, abr. 2017. FapUNIFESP (SciELO). Disponível em: http://www.scielo.br/pdf/tinf/v29n1/0103-3786-tinf-29-01-00091.pdf . Acesso em: 21 abr. 2019.

GONÇALVES, Eduardo Corrêa. **Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas**. 2005. Disponível em: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwiSpN-R4LziAhX0JrkGHUcHBisQFjAAegQIABAC&url=http%3A%2F%2Fwww.dcc.ufla.br%2Finfocomp%2Findex.php%2FINFOCOMP%2Farticle%2Fdownload%2F79%2F64%2F0&usg=AOvVaw1jKgTOt_yup6slyeHVWrc1 . Acesso em: 23 jun. 2019.

HE, Yaobin et al. MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data. **Frontiers Of Computer Science**, [s.l.], v. 8, n. 1, p.83-99, 19 dez. 2013. Disponível em: http://dx.doi.org/10.1007/s11704-013-3158-3. Acesso em: 20 abr. 2019.

HALÉVY, Marc. **A Era do Conhecimento**: princípios e reflexões sobre a revolução noética no século XXI. São Paulo: UNESP, 2010.

HAN, Song; MAO, Huizi; DALLY, Willian j. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. **Computer Vision and Pattern Recognition.** 2016. Disponível em: https://arxiv.org/abs/1510.00149 . Acesso em: 17 mai. 2019.

KUHNEN, Igor Antonio. **Análise de Sistemas de Gerenciamento de banco de dados para armazenamento de dados climáticos**. 2016 . f. 71. Dissertação (Mestre em Física Ambiental) - Universidade Federal do Mato Grosso, Cuiabá, 2016. Disponível em: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwiVg43BhbziAhVEA9QKHSNSBgYQFjAAegQIAhAC&url=http%3A%2F%2Fwww.pgfa.ufmt.br%2Findex.php%2Fbr%2Futilidades%2Fdissertacoes%2F347-igor-antonio-kuhen-disserta%25C3%25A7%25C3%25A3o%2Ffile&usg=AOvVaw2q_BRU5B4gkmbwpGFFHfpG . Acesso em: 20 mai. 2019.

HSSINA, Badr et al. A comparative Study of decision tree ID3 and C4.5. **IJACSA.** p. 13-19, jul 2014. Disponível em: https://saiconference.com/Downloads/SpecialIssueNo10/Paper_3A_comparative_study_of_decision_tree_ID3_and_C4.5.pdf . Acesso em: 29 mar. 2019.

IKHWAN, Ali et al. A Novelty of Data Mining for Promoting Education based on FP-Growth Algorithm. International Journal of Civil Engineering and Technology.

International Journal of Civil Engineering and Technology (IJCIET). Vol. 9 jul/2018 pp.1660-1669. Disponível em: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwje9Kiv8rziAhWwGLkGHSKqCt4QFjAAegQIARAC&url=https%3A%2F%2Fosf.io%2Fjpsfa%2Fdownload%2F%3Fformat%3Dpdf&usg=AOvVaw0S2kyc0hjRD8EuBbHtJn8w . Acesso em: 23 mai. 2019.

ISLER, Cassiano Augusto; PITOMBO, Cira Souza. **Avaliação da escolha modal para o transporte ferroviário de Passageiros na região sudeste através de Árvores de Decisão**. 2014. Disponível em: https://www.revistatransportes.org.br/anpet/article/view/1332/659 . Acesso em: 14 mai. 2019.

JUNKES, Guilherme da Silva. **Evolução da Tecnologia da Informação e Comunicação (TIC) e seus Benefícios para as empresas**. 2014. f.47. Disponível em: http://repositorio.unesc.net/handle/1/2879 . Acesso em: 21 abr. 2019.

KAUR, Noor kamal; KAUR, Usvir; SINGH, Dheerendra. K-medoid Clustering Algorithm - A Review. **International Journal of Computer Application and Techonology (IJCAT)** Vol. 1, Abr/2014. Disponível em: https://www.academia.edu/8446443/K-Medoid_Clustering_Algorithm-A_Review . Acesso em: 22 abr. 2019.

LOPES, Silvia. et al. A Bibliometria e a Avaliação da Produção Científica: indicadores e ferramentas. **Actas do congresso Nacional de bibliotecários, arquivistas e documentalistas,** Lisboa, n. 11, p. 1-7, 2012. Disponível em: https://www.bad.pt/publicacoes/index.php/congressosbad/article/view/429 . Acesso: 13 mar . 2019.

MACHADO JUNIOR, Celso et al. As Leis da Bibliometria em Diferentes Bases de Dados Científicos. **Revista de Ciências da Administração**, Florianópolis, p. 111-123, abr. 2016. ISSN 2175-8077. Disponível em: https://periodicos.ufsc.br/index.php/adm/article/view/2175-8077.2016v18n44p111 . Acesso em: 20 maio 2019.

MARQUES, Luiz Carlos; ULSON, José Alfredo Covolan. A aplicação de redes neurais profundas para detecção e classificação de plantas daninhas e seu estado da arte. **REGRAD - Revista Eletrônica de Graduação do UNIVEM - ISSN 1984-7866**, [S.l.], v. 11, n. 01, p. 391 - 403, aug. 2018. ISSN 1984-7866. Disponível em: https://revista.univem.edu.br/REGRAD/article/view/2638 . Acesso em: 11 apr. 2019.

MEDEIROS, J. M. G. DE; VITORIANO, M. A. V. A evolução da bibliometria e sua interdisciplinaridade na produção científica brasileira. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, v. 13, n. 3, p. 491-503, 25 set. 2015. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8635791 . Acesso em: 10 mai. 2019.

MEDEIROS, Luis Leopardi et al. Three-Parameter Logistic Model (ML3): A

Bibliometrics Analysis. **International Journal of Advanced Engineering Research and Science**. v. 5. p. 128-134. abr. 2018. Disponível em: https://ijaers.com/detail/three-parameter-logistic-model-ml3-a-bibliometrics-analysis/ . Acesso em: 04 mai. 2019.

MELO; Vinícius Veloso. **Clustering de artigos Científicos em uma Ferramenta Inteligente de apoio à Pesquisa.** f. 156. Dissertação (Mestre em Ciência da Computação e Matemática Computacional) - Universidade de São Paulo - São Carlos, 2005. Disponível em: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjPxIyH7LziAhWgILkGHUJyDssQFjAAegQIARAC&url=http%3A%2F%2Fwww.teses.usp.br%2Fteses%2Fdisponiveis%2F55%2F55134%2Ftde-11122014-104427%2Fpublico%2FViniciusVelosodeMelo_ME.pdf&usg=AOvVaw2fMwRmborlNKMa89gKT4zt . Acesso em: 18 mai. 2019.

MONTE, Edson Zanbom; ALBUQUERQUE, T. T. A; REISEN, Valdério Anselmo. Impactos das variáveis meteorológicas na qualidade do ar da região da grande Vitória, Espírito Santo, Brasil, **Rev. bras. meteorol**, São Paulo , v. 31, n. 4, supl. 1, p. 546-554, dec. 2016. Disponível em: http://www.scielo.br/scielo.php?pid=S0102-77862016000500546&script=sci_abstract&tlng=en . Acesso em 20 Mar. 2019.

MOREIRA, Fábio Mosso et al. Tecnologias da Web Semântica para a recuperação de dados agrícolas: um estudo sobre o International Information System of the Agri-cultural Science and Technology (AGRIS). **Revista da Faculdade de Biblioteco-nomia e Comunicação da UFRGS,** Porto Alegre, v.21 n.1 jan/abr. 2015. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/50317/33628 . Acesso em: 21 mai. 2019.

MORISSETTE, Laurence; CHARTIER, Sylvain. The k-means clustering technique: General considerations and implementation in Mathematica. **Tutorials in Quantitative Methods for Psychology**. v.9. pag 15-24. 2013. Disponível em: http://www.tqmp.org/RegularArticles/vol09-1/p015/p015.pdf . Acesso em: 20 mai. 2019.

NANDI, J. C. B et al. **O Algoritmo de Associação Frequent Pattern-Growth na Shell Orion Data Mining Engine**. 2015. Disponível em: http://periodicos.unesc.net/sulcomp/article/view/1787 . Acesso em: 10 abr. 2019.

NELSON, David M. Q. **Uso de redes neurais recorrentes para previsão de séries temporais financeiras**. 2017. 73 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2017. Disponível em: https://www.dcc.ufmg.br/pos/cursos/defesas/1999M.PDF . Acesso em: 22 mai. 2019.

OLIVEIRA, Anderson Francisco. Favorecendo o desempenho do k-means via métodos de inicialização de centróides. **UNIFACCAMP** Campo Limpo Paulista, SP:, 2018. Disponível em: http://www.cc.faccamp.br/Dissertacoes/AndersonFranciscoOliveira.pdf> . Acessoo em: 15 mar. 2019.

OLIVEIRA, Caroline. **Um estudo de caso sobre datasets do Ministério da Justiça: dados brutos ou documentos arquivísticos?**. 2015. 131 f. Tese (Mestrado em Gestão de Documentos e Arquivos) - Universidade Federal do Estado do Rio de Janeiro - Unirio, Rio de Janeiro , 2015. Disponível em: http://www.repositorio-bc.unirio.br:8080/xmlui/bitstream/handle/unirio/11754/Dissertacao_correcao_banca.pdf?sequence=1&isAllowed=y . Acesso em: 11 maio. 2019.

OLIVEIRA, Cristiano Cesar da Silva. **Categorização automática de documentos jurídicos utilizando o classificador naive-bayes**. 2015. 43 f TCC (Tecnólogo em Análise e Desenvolvimento de Sistemas) - IFSP - Campus do Jordão, 2015. Disponível em: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjNmY6i5bziAhVDHrkGHScrCJsQFjAAegQIAxAC&url=http%3A%2F%2Fcristianocesar.com.br%2FBase_controller%2Fdownload%2F1&usg=AOvVaw2b8XHWjPFcz4VsmOZk_nsT . Acesso em. 10 mai. 2019.

OLIVEIRA, Gabriela Silva; PEREIRA, Silvio do Lago. **Implementação de Classificação Bayesiana**. 2017. Disponível em: http://bt.fatecsp.br/system/articles/1036/original/BT%20-%20(Oliveira%20&%20Pereira,%202017)%20%20-%20Implementa%C3%A7%C3%A3o%20de%20Classifica%C3%A7%C3%A3o%20Bayesiana.pdf . Acesso em: 14 abr. 2019.

PRAÇA, Fabíola Silva Garcia. METODOLOGIA DA PESQUISA CIENTÍFICA: ORGANIZAÇÃO ESTRUTURAL E OS DESAFIOS PARA REDIGIR O TRABALHO DE CONCLUSÃO. **Revista Eletrônica "Dialogos Acadêmicos"**, Ribeirão Preto, v. 8, n. 1, p. 72-87, jul. 2015. Disponível em: http://uniesp.edu.br/sites/_biblioteca/revistas/20170627112856.pdf . Acesso em: 31 mai. 2019.

PEREIRA, Fernanda de Carvalho et al. Sistemas de informação e inovação: um estudo bibliométrico. **Journal Of Information Systems And Technology Management**, [s.l.], v. 13, n. 1, p. 81-100, abr. 2016. Disponível em: http://www.scielo.br/pdf/jistm/v13n1/1807-1775-jistm-13-1-0081.pdf . Acesso em: 20 mar. 2019.

POLA, Charles da Luz. **Aplicação de processo de classificação e técnica de Bayes na base de dados de acidentes ocupacionais de uma empresa metalúrgica**. 2018. f 61. Disponível em: https://repositorio.ucs.br/xmlui/bitstream/handle/11338/3913/TCC%20Charles%20da%20Luz%20Pola.pdf?sequence=1&isAllowed=y . Acesso em: 10 abr. 2019.

QUINCOZES, Silvio E. **Detecção de intrusões através da seleção dinâmica de classificador baseado em redes de conselhos.** 2018 . 82 f. Dissertação (Mestre em Ciência da Computação) - Universidade Federal de Santa Maria, Santa Maria, 2018. Disponível em: https://repositorio.ufsm.br/handle/1/14646 . Acesso em: 15 abr. 2019.

RALPHS, Ted. Data Science and Analytics. **Industrial and Systems Engineering Report 15T-009, Lehigh University**. 2015. disponível em: http://coral.ie.lehigh.edu/~ted/files/papers/dataScience.pdf> acesso em: 23 mai. 2019.

REHMAN, Saif Ur; ASGHAR, Sohail; FONG, Simon. DBSCAN: Past, present and future. **He Fifth International Conference on the Applications of Digital Informa tion and Web Technologies**. Fev/2014. Disponível em: https://ieeexplore.ieee.org/abstract/document/6814687/ . Acesso em: 23 mai. 2019.

RIBEIRO, Henrique César Melo; MOLINA, Rodrigo do Carmo; OLIVEIRA, Talmo Curto. Características da Produção Acadêmica sobre Governança Corporativa no Setor Público Divulgadas no Web of Science no Período 1955-2013. **Revista Governança Corporativa.** [São Paulo]. v. 2, n.1 . pp. 94-115, abr. 2015.. Disponível em: https://www.researchgate.net/publication/305056963_Caracteristicas_da_Producao_Academica_sobre_Governanca_Corporativa_no_Setor_Publico_Divulgadas_no_Web_of_Science_no_Periodo_1995-2013 . Acesso em: 02 mai. 2019.

RIEDER, Bernhard. Examinando uma técnica algorítmica: o classificador de bayes como uma leitura interessada da realidade. **Parágrafo**, [S.l.], v. 6, n. 1, p. 123-142, jun. 2018. ISSN 2317-4919. Disponível em: http://revistaseletronicas.fiamfaam.br/index.php/recicofi/article/view/726 . Acesso em: 19 mar. 2019.

RODRIGUES, Fernando; SANT'ANA, Ricardo; FERNEDA, Edberto. Análise do processo de recuperação de conjuntos de dados em repositórios governamentais. **InCID: Revista de Ciência da Informação e Documentação.** 6. 10.11606 issn. 2178-2075.v6i1p38-56 .2015. Disponível em: http://www.revistas.usp.br/incid/article/view/73496 . Acesso em: 20 mai. 2019.

ROQUE, Vitor. MÉTRICAS DA INFORMAÇÃO: o fator de impacto na prática. Egitânia Sciencia. V 10. 2012. Disponível em: https://www.researchgate.net/publication230885956_METRICAS_DA_INFORMACAO_o_fator_de_impacto_na_pratica . Acesso em: 29 mai. 2019.

ROSA, Caroline Silvéria .**Estudo sobre as técnicas e métodos de análise de dados no contexto de Big Data**. 2018. Disponível em: https://repositorio.ufu.br/bitstream/123456789/23710/1/EstudoTecnicasMetodos.pdf . Acesso em. 18 mai. 2019.

RÜEGG, Janine et al. Completing the data life cycle: using information management in macrosystems ecology research. **Frontiers In Ecology And The Environment**, [s.l.], v. 12, n. 1, p.24-30, fev. 2014. Wiley. Disponível em: https://esajournals.onlinelibrary.wiley.com/doi/10.1890/120375 . Acesso em 15 mai. 2019.

SÁ, Alex Guimarães Cardoso de. Evolução automática de algoritmos de redes bayesianas de classificação. Dissertação de Mestrado. Universidade Federal de Minas Gerais. Fev, 2014. Disponível em: https://www.dcc.ufmg.br/pos/cursos/defesas/1726M.PDF . Acesso em: 01 jul. 2019.

SACHET, Marcelo; SILVA, Sheila. **Técnicas de Mineração de Dados Aplicadas na Análise de Dados de Fermentação de Vinhos**. 2018. Disponível em: https://repositorio.ucs.br/xmlui/handle/11338/3810 . Acesso em: 15 mar. 2019.

SÁNCHEZ, María Teresa Ramiro; ALBA-RUIZ, Rubén; RAMIRO, Tamara. Bibliometric profile of RUSC. Universities and Knowledge Society Journal. **Universities and Knowledge Society Journal,** [Granada], v. 11, n. 3, p. 42-62, jul 2014. Disponível em: https://www.researchgate.net/publication/264707137_Bibliometric_profile_of_RUSC_Universities_and_Knowledge_Society_Journal . Acesso: 12 mar . 2019.

SANT'ANA, Ricardo César Gonçalves. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, [S.l.], v. 21, n. 2, p. 116–142, dez. 2016. ISSN 1981-8920. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940. Acesso em: 26mar. 2019.

SANTOS, Alessandro, Ferreira. **Redes Neurais Convolucionais Profundas na Detecção de Plantas Daninhas em Lavoura de Soja**. Tese de Mestrado. Universidade Federal do Mato Grosso do Sul. Mar, 2017. Disponível em: http://www.gpec.ucdb.br/pistori/orientacoes/dissertacoes/alessandro2017.pdf . Acesso em: 27 mar. 2019.

SCHWAB, Klaus.**A Quarta Revolução Industrial**. São Paulo: Edipro, 2016.

SCHUBERT, Erich; ROUSSEEUW, Peter. **Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms**. 2018. Disponível em: https://arxiv.org/abs/1810.05691> . Acessoo em: 13 mar. 2019.

SCHUBERT, Erich; HESS, Sibylle; MORIK, Katharina. The Relationship of DBSCAN to Matrix Factorization and Spectral Clustering. **LWDA**. Vol. 2191. CEUR Workshop Proceedings. 2018, p. 330–334. Disponível em: http://ceur-ws.org/Vol-2191/paper38.pdf . Acesso em: 10 mai. 2019.

SEMELER, Alexandre Ribas. **Ciência da Informação em Contextos de E-Science: bibliotecários de dados em tempos de data science**. 2017. 168 f. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Santa Catarina, Florianópolis, 2017. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/185593/PCIN0168-T.pdf?sequence=-1&isAllowed=y. Acesso em: 10 abr. 2019.

SEHNEM, Simone et al. Rede de Cooperação entre Autores que Publicam nas Temáticas Stakeholders, Agro e Bioenergia, Biocombustíveis e Sustentabilidade.

Desenvolvimento em Questão. [S.l] p. 289-315, 2013. Disponível em: https://www.researchgate.net/publication/315905725_Rede_de_Cooperacao_entre_Autores_que_Publicam_nas_Tematicas_Stakeholders_Agro_e_Bioenergia_Biocombustiveis_e_Sustentabilidade/download . Acesso em: 02 jun. 2019.

SHCHERBAKOV, Maxim et al. Lean Data Science Research Life Cycle: A Concept for Data Analysis Software Development. **Communications in Computer and Information Science**.[SI] p.708-716, 2014. Disponível em: https://www.researchgate.net/publication/278709585_Lean_Data_Science_Research_Life_Cycle_A_Concept_for_Data_Analysis_Software_Development . Acesso em: 10 mai. 2019.

SIDHU, Shivam et al. (2014). FP Growth Algorithm Implementation. International Journal of Computer Applications. **International Journal of Computer Applications,** maio. 2014. Disponível em: https://pdfs.semanticscholar.org/ef40/f70587ba1c653389c08cd21e96e387c5e6b8.pdf . Acesso em: 23 mai. 2019.

SILVA, Alcione Dias et al. Mineração de dados aplicada a relação de clientes e pagamentos - Estudo Bibliométrico. **Persp. online: exatas & engenharia**. Campos dos Goytacazes. v.3, n.5, p.45-59. 2013. Disponível em: http://www.seer.perspectivasonline.com.br/index.php/exatas_e_engenharia/article/view/88 . Acesso em: 02 mai. 2019.

SONG, Yan-yan; LU, Ying. Decision tree methods: applications for classification, and prediction. **Biostatistics in psichiatry**. v.27, n .2, p.130-135. 2015. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/ . Acesso em: 05 mai. 2019.

SOUSA, Saymon Ricardo de Oliveira et al. Análise de Correlação e Regressão como Ferramenta para Gestão da Manutenção: um Estudo Aplicado na Indústria de Mineração e Logística. **Revista Fsa**, [s.l.], v. 15, n. 6, p.151-167, 1 nov. 2018. Revista FSA. Disponível em : http://dx.doi.org/10.12819/2018.15.6.8. Acesso em: 04 mai. 2019.

STEPHANIE C. Hicks; RAFAEL A. Irizarry. A Guide to Teaching Data Science. 2017 disponível em: https://arxiv.org/abs/1612.07140. Acesso em: 23 mai. 2019.

TAKECIAN, Pedro Losco. **Diretrizes metodológicas e validação estatística de dados para a construção de data warehouses.** 2014. 112 f. Tese (Doutorado em Ciências) - Instituto de Matemática e Estatśstica de São Paulo da Universidade de São Paulo, São Paulo, 2014. Disponível em: http://www.teses.usp.br/teses/disponiveis/45/45134/tde-10112014-110134/pt-br.php . Acesso em. 22 mai. 2019.

VASCONCELOS, Lívia Maria Rocha de, CARVALHO Cedric Luiz de. Aplicação de Regras de Associação para Mineração de Dados na Web, v. 1 n. 1 (2018): **Revista Telfract** - n º 1/2018. Disponível em:

http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf . Acesso: 16 mar. 2019.

WANG, Shusen; GITTENS, Alex; MAHONEY, Michael W. Scalable Kernel K-Means Clustering with Nyström Approximation: Relative-Error Bounds. **Journal of Machine Learning Research**. nº 20 pag 1-49. fev/2019. Disponível em: https://arxiv.org/abs/1706.02803 . Acesso em: 18 mai. 2019.

WEBER, Julia Silva. **Eliminação Segura de Arquivos em Memória Não-Volátil**. 2017. 108 f. Dissertação (Mestre em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre 2017. Disponível em: http://tede2.pucrs.br/tede2/bitstream/tede/7546/2/DIS_JULIA_SILVA_WEBER_COMPLETO.pdf . Acesso em: 17 mai. 2019.

ZUEGE, Tiago Jasper. **Aplicação de técnicas de Mineração De Dados para detecção de perdas comerciais na distribuição de energia elétrica**. 2018. Disponível em: https://www.univates.br/bdu/bitstream/10737/2240/1/2018TiagoZuege.pdf . Acesso em: 20 mai. 2019.

APÊNDICE

APÊNDICE A – RELAÇÃO DE ARTIGOS CONTIDOS NA FASE FINAL DA BIBLIOMETRIA

1 - A Review of DAN2 (Dynamic Architecture for Artificial Neural Networks) Model in Time Series Forecasting1.
Disponível em: http://www.scielo.org.co/pdf/inun/v16n1/v16n1a08.pdf
2 - FICOBU: Filipino WordNet construction using decision tree and language modeling.
Disponível em: http://www.ijmlc.org/vol9/772-ML0001.pdf
3 - Recursive Hierarchical Clustering Algorithm.
Disponível em: http://www.ijmlc.org/vol8/654-L0114.pdf
4 - Strategy Representation by Decision Trees in Reactive Synthesis.
Disponível em: https://arxiv.org/pdf/1802.00758.pdf.
5 - Interactive Visual Decision tree for developing detection rules of attacks on web applications.
Disponível em: https://thesai.org/Downloads/Volume9No7/Paper_5Interactive_Visual_Decision_Tree.pdf
6 - Using game theory to handle missing data at prediction time of ID3 and C4.5 algorithms.
Disponível em : https://thesai.org/Downloads/Volume9No12/Paper_32-Using_game_theory_to_handle_missing_data.pdf
7- Machine learning-based malicious application detection of android.
Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8101455
8 - HC-CART: A parallel system implementation of data mining classification and regression tree (CART) algorithm on a multi-FPGA system.
Disponível em: http://delivery.acm.org/10.1145/2410000/2400706/a47-chrysos.pdf?ip=179.216.20.231&id=2400706&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1562114705_0ff36d9ef34641c1e587df984b8aeb1c

9 - Training of feedforward neural networks for data classification using hybrid particle swarm optimization, Mantegna Lévy flight and neighborhood search. Disponível em: https://www.sciencedirect.com/science/article/pii/S2405844018367161

10 - Robust Exponential Memory in Hopfield Networks. Disponível em: https://link.springer.com/content/pdf/10.1186%2Fs13408-017-0056-2.pdf

11 - An approach for automating the design of convolutional neural networks. Disponível em : https://iopscience.iop.org/article/10.1088/1757-899X/450/5/052004/pdf

12 - New 3D objects retrieval approach using multi agent systems and artificial neural network. Disponível em: https://thesai.org/Downloads/Volume9No12/Paper_57-New_3D_Objects_Retrieval_Approach.pdf

13 - A two-step neural dialog state tracker for task-oriented dialog processing. Disponível em : downloads.hindawi.com/journals/cin/2018/5798684.pdf

14 - Quantum entanglement in neural network states. Disponível em : https://arxiv.org/pdf/1701.04844.pdf

15 - WSN-DS: A Dataset for Intrusion Detection Systems in Wireless Sensor Networks. Disponível em : downloads.hindawi.com/journals/js/2016/4731953.pdf.

16 - Automatic Construction and Global Optimization of a Multisentiment Lexicon. Disponível em : downloads.hindawi.com/journals/cin/2016/2093406.pdf

17 - Improving Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts. Disponível em: downloads.hindawi.com/journals/cin/2016/1638936.pdf

18 - Operations on quantum physical artificial neural structures. Disponível em : https://www.sciencedirect.com/science/article/pii/S1877705814003944.

19 - Self-Learning facial emotional feature selection based on rough set theory. Disponível em: downloads.hindawi.com/journals/mpe/2009/802932.pdf

20 - Connectionist modal logic: Representing modalities in neural networks.
Disponível em:
https://www.sciencedirect.com/science/article/pii/S030439750600750X
21 - Local feature extraction from RGB and depth videos for human action recognition.
Disponível em: http://www.ijmlc.org/vol8/699-V0050.pdf.
22 - A Multiple Kernel Learning Model Based on p -Norm.
Disponível em: downloads.hindawi.com/journals/cin/2018/1018789.pdf
23 - Combination of long-term and short-term features for age identification from voice.
Disponível em : www.aece.ro/displaypdf.php?year=2018&number=2&article=13
24 - Triangle shape feature based on selected centroid for Arabic subword handwriting.
Disponível em :
https://pdfs.semanticscholar.org/6f53/69249e39d8897cc3902e61d1156ac7c0e2dd.pdf
25 - WordNet based implicit aspect sentiment analysis for crime identification from Twitter.
Disponível em : https://thesai.org/Downloads/Volume9No12/Paper_22-WordNet_based_Implicit_Aspect_Sentiment_Analysis.pdf
26 - An Extension of the VSM Documents Representation using Word Embedding.
Disponível em : https://content.sciendo.com/downloadpdf/journals/cplbu/3/1/article-p249.xml
27 - Speech query recognition for Tamil language using Wavelet and Wavelet Packets.
Disponível em: jips-k.org/file/down?pn=80070
28 - Iris image recognition based on independent component analysis and support vector machine.
Disponível em :
https://www.researchgate.net/publication/282559804_Iris_Image_Recognition_Based_on_Independent_Component_Analysis_and_Support_Vector_Machine/download

29 - Application of flexible logic average operation model on selection of approximate support vectors on [0,8) interval.
Disponível em : http://docsdrive.com/pdfs/ansinet/itj/2013/5377-5385.pdf
30 - Segmentation strategy of handwritten connected digits (SSHCD).
Disponível em : https://link.springer.com/content/pdf/10.1007%2F978-3-642-24088-1_26.pdf
31 - Support vector machine regression algorithm based on chunking incremental learning.
Disponível em: https://www.researchgate.net/publication/225789609_Support_Vector_Machine_Regression_Algorithm_Based_on_Chunking_Incremental_Learning/download.
32 - A new DDoS detection model using multiple SVMs and TRA.
Disponível em : https://link.springer.com/content/pdf/10.1007%2F11596042_100.pdf
33 - Incomplete Multiview Clustering via Late Fusion.
Disponível em: downloads.hindawi.com/journals/cin/2018/6148456.pdf
34 - A simple density with distance based initial seed selection technique for K means algorithm.
Disponível em: http://cit.fer.hr/index.php/CIT/article/view/3605/2200
35 - Consensus kernel K-means clustering for incomplete multiview data.
Disponível em : downloads.hindawi.com/journals/cin/2017/3961718.pdf
36 - K-means algorithm with a novel distance measure.
Disponível em: https://www.researchgate.net/publication/259496474_K-Means_Algorithm_with_a_Novel_Distance_Measure/download
37 - A bad instance for k-means++.
Disponível em : https://www.sciencedirect.com/science/article/pii/S0304397512001806
38 - An adaptive initial cluster center selection k-means algorithm and implementation.
Disponível em : http://docsdrive.com/pdfs/ansinet/itj/2013/5665-5668.pdf
39 - The planar k-means problem is NP-hard.
Disponível em:

https://www.sciencedirect.com/science/article/pii/S0304397510003269

40 - Multi-objective optimization for adaptive web site generation.

Disponível em: https://link.springer.com/content/pdf/10.1007%2F11590316_105.pdf

41 - A new clustering approach to identify the values to query the deep web access forms.

Disponível em: https://ieeexplore.ieee.org/document/8398666

42 - G-DBSCAN: A GPU accelerated algorithm for density-based clustering.

Disponível em :

https://www.sciencedirect.com/science/article/pii/S1877050913003438

43 - Frequent itemset mining in big data with effective single scan algorithms.

Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8529189

44 - Heuristic Frequent Term-Based Clustering of News Headlines.

Disponível em :

https://www.sciencedirect.com/science/article/pii/S221201731200597X

45 - Integrating Signature Apriori based Network Intrusion Detection System (NIDS) in Cloud Computing.

Disponível em :

https://www.sciencedirect.com/science/article/pii/S221201731200655X

46 - Performance of Distributed Apriori Algorithms on a Computational Grid.

Disponível em: https://ieeexplore.ieee.org/document/5394128.

APÊNDICE B – ANÁLISES IMPLÍCITAS DE DADOS NA PRODUÇÃO DO CONHECIMENTO EM CIÊNCIA DA COMPUTAÇÃO: UM ESTUDO BIBLIOMÉTRICO

Análises Implícitas de Dados na Produção do Conhecimento em Ciência da Computação: um Estudo Bibliométrico

Diovan O. Leal[1], Merisandra C. Mattos[2], Kristian Madeira[3]

[123]Ciência da Computação – Universidade do Extremo Sul Catarinense (UNESC)

Caixa Postal 3.167 – 88.806-000 – Criciúma – SC – Brazil

{diovan}@hotmail.com, {mem,kristian}@unesc.net

Abstract. *This work aims to develop a bibliometric research in computer science from works that employ techniques of implicit analysis. Besides to the bibliometric mapping, the theoretical basis on implicit analysis and bibliometry was also carried out. The following implicit analyzes are addressed: Apriori, decision trees, Bayesian classifiers, DBSCAN, FP-Growth, support vector machines, artificial neural networks, k-means and k-medoid. The scientific article analyzed come from three databases, SciElo, Scopus and Web of Science.*

Resumo. Este trabalho tem por objetivo, desenvolver uma pesquisa bibliométrica na ciência da computação a partir de trabalhos que empregam técnicas de análises implícitas. Além do mapeamento bibliométrico, também foi realizada a fundamentação teórica sobre, análises implícitas e bibliometria. São abordadas as seguintes análises implícitas: Apriori, árvores de decisão, classificadores *bayesianos*, *DBSCAN, FP-Growth,* máquinas de vetores de suporte, redes neurais artificiais, *k-means e k-medoid*. Os artigos científicos analisados são oriundos de três bases de dados, *SciElo*, *Scopus* e *Web of Science*.

1. Introdução

Nas produções científicas no campo de ciência orientada a dados, observa-se a predominância de um objetivo macro comum, a capacidade de extrair conhecimento de grandes e heterogêneas bases de dados, por meio de métodos de análises implícitas [Bufrem et. al 2016].

As pesquisas nesta área de conhecimento aplicam estes métodos de análise de dados e considerando a importância do assunto é interesse científico que estudos de mapeamento da ciência auxiliem no mapeamento das produções científicas.

Neste sentido, o emprego da bibliometria pode resolver importantes questões em torno

de informações úteis aos pesquisadores referentes a produção de conhecimento na área de dados dentro da computação.

Portanto, este trabalho tem por objetivo, mapear por meio da bibliometria a produção científica deste cenário de pesquisa.

2. Análises Implícitas

No contexto de ciência dos dados, onde volume, variedade e velocidade são aspectos pertinentes a este campo de estudo, é fato que métodos específicos de análise de dados sejam empregados para a produção do conhecimento, estes métodos são as análises implícitas [Pola 2018], [Rosa 2018].

As análises implícitas são aplicadas em casos onde a informação não está explícita e de fácil acesso. Quando a pergunta envolve dados que não estão estruturados, não possuem relação direta ainda possuem fontes e formatos variados e são volumosos. Para estes casos os métodos de análises implícitos são úteis e necessários [Amaral 2016].

Desta forma, de acordo com o autor Fernando Amaral em que na sua obra: Introdução à Ciência de Dados: mineração de dados e big data, aborda as análises implícitas assim como traz exemplos destas técnicas de análise.

De forma geral é exposto a definição dos seguintes métodos que serão investigados na pesquisa bibliométrica: árvores de decisão, classificadores *bayesianos*, redes neurais artificiais, máquina de vetores de suporte, *k-means*, *k-medoid*, *DBSCAN* , Apriori e *FP-Growth* [Amaral 2016].

As árvores de decisão são uma das técnicas para representar conhecimento, os conjuntos de dados a serem treinados são iterativamente subdivididos em subconjuntos até cada subconjunto pertencer a uma mesma classe ou uma classe se destacar cessando então as divisões, resultando em dados organizados e compactos [Sachet e Silva 2018].

Classificadores *bayesianos* são uma classe de algoritmos que fornecem métodos para inferências estatísticas na classificação de elementos [Rieder 2018].

As redes neurais artificiais são um modelo de processamento em forma de rede inspirado na biologia, cada nó da rede representa um neurônio artificial e processa determinados cálculos, gerando uma resposta única [Binoti, D. Binoti e Leite 2014].

Máquinas de vetores de suporte são classificadores de dados fundamentadas em teoria de aprendizado estatístico de aplicação genérica [Dosciatti, Paterno e Paraiso 2013].

K-means e *k-medoid* são algoritmos particionais não hierárquicos utilizados para resolver problemas de agrupamento [Brito J, Semaan e Brito L 2014], [Oliveira 2018], [Schubert e Rousseeuw 2018].

Density Based Spatial Clustering of Application with Noise (*DBSCAN*) é um método de *clusterização* não paramétrico baseado em densidade [He et al 2014].

Apriori é um algoritmo de regra de associação utilizado em mineração de dados, que

tem como objetivo identificar frequências de conjuntos de itens em transações e construir regras a partir destes elementos [Elgaml et al 2015].

FP-Growth também é um algoritmo para regras de associação, implementa conceito de dividir para conquistar e uma estrutura de árvore conhecida como *Frequent Pattern Tree* (*FP-tree*) o que evita a necessidade de escanear a base inteira repetidas vezes [Nandi et al 2015], [Sidhu et al 2014].

Estes métodos e algoritmos apresentados, frequentemente aparecem em artigos científicos que tratam problemas acerca de dados, sendo assim esta pesquisa bibliométrica os investigará.

3. Bibliometria

A bibliometria é uma área de estudos do campo da ciência da informação, que de forma quantitativa e estatística avalia as publicações científicas e decorrente disso gera indicadores relativos à produção de conhecimento, auxiliando no entendimento e direcionamento das pesquisas. [Medeiros e Vitoriano 2015], [Sánchez, Alba-Ruiz e Ramiro 2014, tradução nossa].

Quanto à fundamentação, a bibliometria é regida por três leis principais, lei de Bradford, Zipf e Lotka.

Bradford aferi quão reputável um periódico é, possibilitando detectar os periódicos mais importantes e que prioriza a circulação de determinado tema em específico. Ainda, os artigos estão distribuídos igualmente em três zonas de periódicos, muito produtivas, produtividade intermediária e pouco produtivas [Chueke e Amatucci 2015], [Machado Junior et. al 2016].

Zipf quantifica a frequência de palavras-chave, revelando os temas mais recorrentes em uma linha de pesquisa, resultando uma lista ordenada de termos mais frequentes por assunto [Ribeiro, Molina e Oliveira 2015].

Lotka mensura a produtividade dos autores, identificando os mais produtivos em um campo do conhecimento, propondo que a maioria das publicações advêm da minoria dos autores. [Chueke e Amatucci 2015], [Machado Junior et. al 2016].

A bibliometria e suas leis possibilitam a avaliação da produção científica e isto ocorre por meio de grupos de indicadores bibliométricos de quantidade e qualidade científica.

Qualidade científica é um índice obtido por meio da crítica realizada ao conteúdo de uma publicação, enquanto que atividade científica é uma medida, e portanto quantitativa acerca do, número de trabalhos publicados, número da produção individual dos autores, número de coautorias entre outros [Lopes et. al 2012].

Quanto à produtividade e impacto dos pesquisadores, é utilizado o indicador *h-index,* este índice trabalha com o número das citações dos artigos, lista-se todos os artigos de um autor ordenando-a decrescentemente por número de citações, quando existe um número

máximo de citação que contempla a todas as citações então este é o *h-index* do autor.

A pesquisa bibliométrica além das leis e indicadores conta com o auxílio de ferramentas para o desenvolvimento da investigação, as ferramentas de pesquisa bibliométrica em maior uso são *Web of Science, Scopus e Google Scholar Metrics,* por meio destas fontes o investigador científico tem a sua disposição para consulta indicadores como o *h-index*, assim como inúmeros artigos de pesquisa. [Lopes et. al 2012].

4. Metodologia

Este trabalho, é uma pesquisa bibliométrica direcionada a investigar a produção de conhecimento em ciência da computação pelo emprego das análises implícitas, logo a bibliometria é utilizada como referencial metodológico. Portanto trata-se de uma pesquisa quantitativa visto que tem como objetivo, medir, aferir e quantificar.

Para o desenvolvimento desta exploração científica foram definidas determinadas etapas de execução. Escolha das análises implícitas, área de concentração da aplicação das análises, definição da estratégia de busca, definição das bibliotecas eletrônicas, pesquisa exploratória, download e organização dos artigos, análise superficial dos artigos, análise detalhada dos artigos, extração de dados e análise e relacionamento dos dados levantados.

A primeira etapa define quais análises implícitas compõe a pesquisa. De acordo com o autor Fernando Amaral em que na sua obra: Introdução à Ciência de Dados: mineração de dados e big data, aborda as análises implícitas, portanto, as seguintes análises serão investigadas na bibliometria: árvores de decisão, classificadores *bayesianos*, redes neurais artificiais, máquina de vetores de suporte, *k-means*, *k-medoid*, *DBSCAN*, Apriori e *FP-Growth.* Quanto ao campo de atuação, os artigos devem resolver problemas na esfera da computação [Amaral 2016].

Com as análises e o campo já delimitados foram criadas as estratégias de busca, cada estratégia ficou estruturada com o nome da análise em inglês entre aspas duplas, seguido do operador lógico *AND* mais o nome da área em inglês entre aspas duplas.

Portanto estas foram as estratégias empregadas, *"Decision trees" AND "Computer science", "Bayesian Classifiers" AND "Computer science", "Bayesian Classification" AND "Computer science", "Artificial neural networks" AND "Computer science", "Support Vector Machine" AND "Computer science", "K-means" AND "Computer science","K-medoid" AND "Computer science", "DBSCAN" AND "Computer science", "Apriori" AND "Computer Science"* e *"FP-Growth" AND "Computer Science".*

A próxima etapa é a escolha das bibliotecas digitais que foram pesquisadas, quanto a isto, por serem mais difundidas no meio científico, as seguintes fontes de dados foram utilizadas: *SciElo*, *Scopu*s e *Web of Science*.

Nestas bases de dados foram aplicados os seguintes critérios de buscas, além da *string* de consulta, o único filtro aplicado é referente à artigos que sejam de acesso público, ou seja, sem custos ou restrições, também não foi restringido por data, autor ou qualquer outro limitador de busca.

A pesquisa exploratória foi realizada nas datas compreendidas entre 08 e 11 de abril de 2019. Os resultados de quantidades de artigos *X* estratégia *X* base de dados foram tabulados e serviram como base para a próxima etapa da pesquisa.

A etapa de download e organização dos arquivos foi desenvolvida entre os dias 25 de abril e 03 de maio de 2019.

Após fazer o *download* e organizar os arquivos foi desenvolvida uma tabela contendo as seguintes colunas, "periódico", "análise", "título do artigo", "sim/não", "aplicação" e "observação". Então foi realizada a leitura superficial de cada artigo, que consiste da leitura do título, subtítulo, resumo, palavras-chave e título dos capítulos, feito isto os artigos foram tabulados com os seguintes dados, nomes dos periódicos, nomes dos artigos, o campo sim ou não refere-se ao enquadramento do artigo com nosso objetivo de pesquisa, o campo aplicação é referente a área de estudo que é identificada no artigo e o campo observação para qualquer informação adicional.

Quanto ao enquadramento do artigo, foi usado como critério estar evidente o uso de uma das análises na área da ciência da computação e o documento deve ser um artigo, livros ou artigos estritamente caracterizados como levantamento bibliográfico ou análises bibliométricas foram descartados para a próxima fase.

Ao fim da primeira triagem dos artigos, é iniciada a etapa de análise detalhada dos arquivos, que consiste da leitura completa destes.

Novamente, os registros que possuem o valor *sim* na coluna sim/não e são aplicados à computação foram mantidos para próxima fase, estes foram copiados para nova aba da tabela, que contém as seguintes colunas, "nº", "autor", "*h-index"*, "coautores", "ano", "universidade", "país", "revista", "*qualis"*, "título", "palavras-chave", "objetivo", "metodologia", "resultados", "limitação", "conclusão", "análise empregada" e "subcampo de aplicação".

Nesta etapa algumas informações foram transcritas do artigo para a planilha, como: autor, coautores, ano, universidade, país, revista, título e palavras-chave. Outras necessitaram análise e extração de informações como: objetivo, metodologia, resultados, limitação, conclusão e técnica empregada. Enquanto outras precisaram ser consultadas em fontes específicas como *h-index*, que é consultado na base da Scopus, assim como o *qualis* que é consultado na plataforma Sucupira.

A última planilha representa os dados finais da pesquisa, a partir destes foi realizada a análise bibliométrica, aplicando as leis basilares da bibliometria, e índices bibliométricos pertinentes, possibilitando então, o entendimento das tendências das pesquisas da área assim como lacunas.

5. Resultados

Nesta seção os resultados serão apresentados e discutidos em relação às leis bibliométricas.

Está pesquisa na primeira fase contou com duzentos e oitenta e três artigos. Após as

etapas de análise a quantidade de artigos foi reduzida para quarenta e seis. Destacam-se: base *Scopus* que indexa quarenta e dois artigos, as duas análises com maior quantidade de artigos: *artificial neural networks* com treze artigos e *support vector machine* com doze artigos e ausência de artigos com a análise *bayesian classification* ou *classifiers*.

Quanto aos autores, são quarenta e cinco, apenas um autor tem dois artigos presentes nesta fase, o pesquisador Chinês Ye, Yongkai, filiado a *National University of Defense Technology* , seus dois trabalhos empregam a técnica de análise implícita *k-means.*

Referente ao *h-index,* os autores com maiores h-index são: Brázdil Thomáš, Artur S. D'Avila Garcez e Mahajan, Meena com os respectivos *h-index* quinze, treze e doze. O *h-index* é um indicador de impacto e produtividade do autor, por inferência o resultado obtido confirma a lei de Lotka, que propõe que a maioria das publicações advêm da minoria dos autores [Machado Junior et. al 2016].

Ainda, analisando o cenário de autoria e coautoria, entre os coautores, apenas Xinwang Liu, Qiang Liu e Jianping Yin participam em mais de um trabalho, analisando este dado em específico, foi identificado que estes três autores publicam juntamente com o autor Ye, Yongkai, e, estes trabalhos têm como foco a análise *k-means.*

Quanto a produtividade por ano, 2018 tem a maior quantidade de artigos, dezesseis, nos demais anos a produtividade não é tão acentuada quanto a dos anos de 2013 e 2017, seis artigos em cada ano, já em 2012 tem quatro artigos. Observa-se que a produtividade anual anterior a 2018 tem leves oscilações no entanto o ano de 2018 apresenta notável crescimento.

Analisando a produtividade por país, destacam-se países como: China e Índia, que mais contribuíram com a pesquisa, nove e sete artigos respectivamente, a maioria absoluta dos demais países produziu um artigo. Novamente, por inferência o princípio de Lotka é observado, poucos produzindo muito e muitos com pouca produção.

Quanto aos periódicos, os mesmos são avaliados por quantidade e qualidade. Quantitativamente destacam-se os periódicos *Hindawi, International Journal of Advanced Computer Science and Applications* e *IEE access* com oito, cinco e quatro artigos respectivamente. Qualitativamente tendo como referência o *qualis*, destaca-se o periódico *Theoretical Computer Science* com avaliação máxima, A1. Ainda, observa-se que muitas revistas não possuem esta avaliação *qualis* na plataforma Sucupira.

Estabelecendo uma relação entre revistas e artigos. É possível confirmar alguma proximidade com a lei de Bradford, que propõe zonas de produtividade. Nesta pesquisa a zona muito produtiva é composta por duas revistas com cinco e oito artigos cada, a zona com produtividade razoável é formada por quatro revistas que possuem quatro ou três artigos e finalmente a zona menos produtiva com dezenove periódicos com um ou dois artigos, estabelecendo assim uma relação com o que Bradford afirma [Alvarado 2007].

Os artigos também foram analisados quanto as palavras-chave. Nesta pesquisa foram catalogadas as palavras-chave que estão explicitamente declaradas. Das palavras catalogadas destaca-se *clustering, k-means, support vector machine, computing, network, mining, system, tree, artificial neural network e classification.*

Conforme análise dos artigos, pode-se identificar grupos de artigos com objetivos semelhantes, foi identificado cinco grupos nos quais os artigos se enquadram.

Grupo de artigos que de forma geral objetivam resolver algum problema relacionado a performance de algoritmo, ou pesquisa de aperfeiçoamento de desempenho em arquitetura computacional ou a correção de algum processo de algoritmo ou arquitetural da computação. É entendido como: pesquisa e desenvolvimento. Estas características estão presentes em vinte e seis artigos.

Quanto a análise adotada por este grupo, são abordadas pelo menos uma destas técnicas, RNA, árvores de decisão, *k-means, SVM, DBSCAN,* Apriori e Fp-growth.

Há também trabalhos voltados para o Processamento de Linguagem Natural, estes trabalhos tem como foco desenvolver tecnologias que permitam que a interação homem-máquina seja mais transparente e natural, possibilitando às máquinas reconhecer as mais diversas formas pelas quais os seres humanos, se comunicam e se expressão, doze artigos compõem este grupo, e as análises utilizadas são árvores de decisão, RNA e SVM.

A segurança computacional também é abordada por alguns artigos, cinco artigos estão alinhados com este objetivo de pesquisa e utilizam árvores de decisão, RNA e Apriori.

Detectado que um artigo tem como objetivo contornar problemas relacionados as lacunas de dados em *datasets*, considerado um problema para as análises de dados e utiliza árvores de decisão para propor soluções neste campo.

Finalmente o quinto grupo está relacionado à pesquisa e indexação de conteúdo, dois artigos têm suas pesquisas voltadas a este propósito e utilizam *DBSCAN* e *k-medoid* nas soluções propostas.

Pela perspectiva de grupos de pesquisa, as árvores decisão são empregadas em quatro dos cinco grupos, sendo a análise mais utilizada, Apriori e RNA são empregadas em três grupos, *DBSCAN* é utilizada em dois grupos e é utilizada em apenas um grupo as análises *k-means, k-medoid, SVM e FP-Growth.*

Finalmente, o último indicador bibliométrico analisado é quanto ao idioma das pesquisas, constatou-se que todos os artigos desta pesquisa estão na língua inglesa.

6. Conclusão

Esta pesquisa por meio da bibliometria mapeou a produção de conhecimento em ciência da computação, no escopo de artigos que empregam análises implícitas para resolver problemas específicos da área. De acordo com a pesquisa, neste cenário inglês é a língua dominante. Ainda segundo análise de produtividade anual entende-se que este campo de pesquisa encontra-se em evolução. Quanto aos autores destacam-se: Ye Yongkai, Brázdil Thomáš, Artur S. D'Avila Garcez e Mahajan, Meena. Também identificado dois países mais produtivos: China e Índia. Relacionado às análises, as mais utilizadas são: árvores de decisão, redes neurais artificiais e *Apriori.* Identificado cinco campos de pesquisa, entre eles duas possíveis tendências e uma possível lacuna, além das palavras -chave mais utilizadas.

Referências

Alvarado, R. U. (2007) "A Bibliometria: história, legitimação e estrutura". In Para entender a ciência da informação (Salvador EDUFBA). Disponível em: https://www.academia.edu/1390400/A_BIBLIOMETRIA_HISTORIA_LEGITIMA%C3%87%C3%83O_E_ESTRUTURA . Acesso em: 05 mai. 2019.

Amaral, Fernando. (2016) "Introdução à ciência de dados: mineração de dados e big data". Rio de Janeiro: Alta Books.

Binoti, D., Leite, H. e Binoti, M. (2014) "Configuração de Redes Neurais Artificiais para estimação do volume de árvores". In Ciência da Madeira (Braz. J. Wood Sci) (Pelotas, v.05, n. 01, p. 58-67, Maio 2014 ISSN: 2177 - 6830). Disponível em: https://periodicos.ufpel.edu.br/ojs2/index.php/cienciadamadeira/index . Acesso em: 17 mar. 2019.

Brito, J. A. M., Semaan, G. S., Brito, L. R. (2014) "Resolução do problema dos k- medoids via algoritmo genético de chaves aleatórias viciadas". In Anais do Xvii Simpósio de Pesquisa Operacional e Logística da Marinha ([s.l.], p.50-61, ago. 2014. Editora Edgard Blücher). Disponível em: https://www.researchgate.netpublication269203283_RESOLUCAO_DO_PROBLEMA_DOS_KMEDOIDS_VIA_ALGORITMO_GENETICO_DE_CHAVES_ALEATORIAS_VICIADAS . Acesso em: 05 mai. 2019.

Bufrem, L S et al. (2016) "Produção Internacional Sobre Ciência Orientada a Dados: análise dos termos Data Science e E-Science na Scopus e na Web of Science". In Informação & Informação (Londrina, v. 21, n. 2, p.40-67, dez. 2016. Universidade Estadual de Londrina). Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/26543/20114 . Acesso em : 21 Abr. 2019.

Chueke, G. V., Amatucci. (2015) "O que é bibliometria? Uma introdução ao Fórum". In Revista Eletrônica de Negócios Internacionais (São Paulo, v. 10, n. 2, p. 1-5, mai./ago. 2015). Disponível em: http://internext.espm.br/internext/article/view/330 . Acesso em: 21 mar. 2019.

Dosciatti, M., Paterno, L. e Paraiso, E. (2013) "Identificando Emoções em Textos em Português do Brasil usando Máquina de Vetores de Suporte em Solução Multiclasse". Disponível em: https://www.researchgate.net/publication/277813389_Identificando_Emocoes_em_Textos_em_Portugues_do_Brasil_usando_Maquina_de_Vetores_de_Suporte_em_Solucao_Multiclasse . Acesso em: 15 mar. 2019.

Elgaml, E., et al. (2015) "Improved FP-Growth Algorithm with Multiple Minimum

Supports Using Maximum Constraints". In World Academy of Science, Engineering and Technology, International Science Index 101 (Dubai 2015) Disponível em: https://www.researchgate.net/publication/280611828_Improved_FP-Growth_Algorithm_with_Multiple_Minimum_Supports_Using_Maximum_Constraints/download . Acesso em: 20 abr. 2019.

He, Y., et al. (2013) "MR-DBSCAN: a scalable MapReduce-based DBSCAN algorithm for heavily skewed data". In Frontiers Of Computer Science ([s.l.], v. 8, n. 1, p.83-99, 19 dez. 2013) Disponível em: http://dx.doi.org/10.1007/s11704-013-3158-3. Acesso em: 20 abr. 2019.

Lopes, S. et al (2012) "A Bibliometria e a Avaliação da Produção Científica: indicadores e ferramentas" In Actas do congresso Nacional de bibliotecários, arquivistas e documentalistas (Lisboa, n. 11, p. 1-7, 2012) Disponível em: https://www.bad.pt/publicacoes/index.php/congressosbad/article/view/429 . Acesso: 13 mar . 2019.

Machado J, C et al (2016) "As Leis da Bibliometria em Diferentes Bases de Dados Científicos". In Revista de Ciências da Administração (Florianópolis, p. 111-123, abr. 2016. ISSN 2175-8077) Disponível em: https://periodicos.ufsc.br/index.php/adm/article/view/2175-8077.2016v18n44p111. Acesso em: 20 maio 2019.

Medeiros, J. M. G e Vitoriano, M. A. V. (2015) "A evolução da bibliometria e sua interdisciplinaridade na produção científica brasileira" In RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação (v. 13, n. 3, p. 491-503, 25 set. 2015). Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8635791 . Acesso em: 10 mai. 2019.

Nandi, J. C. B et al. (2015) "O Algoritmo de Associação Frequent Pattern-Growth na Shell Orion Data Mining Engine". Disponível em: http://periodicos.unesc.net/sulcomp/article/view/1787 . Acesso em: 10 abr. 2019.

Oliveira, A. F. (2018) "Favorecendo o desempenho do k-means via métodos de inicialização de centróides". In UNIFACCAMP (Campo Limpo Paulista, SP:, 2018) Disponível em: http://www.cc.faccamp.br/Dissertacoes/AndersonFranciscoOliveira.pdf> . Acessoo em: 15 mar. 2019.

Pola, C. L. (2018) "Aplicação de processo de classificação e técnica de Bayes na base de dados de acidentes ocupacionais de uma empresa metalúrgica". Disponível em: https://repositorio.ucs.br/xmlui/bitstream/handle/11338/3913/TCC%20Charles%20da%20Luz%20Pola.pdf?sequence=1&isAllowed=y . Acesso em: 10 abr. 2019.

Ribeiro, H. C. M., Molina, R. C. e Oliveira, T. C. (2015) “Características da Produção Acadêmica sobre Governança Corporativa no Setor Público Divulgadas no Web of Science no Período 1955-2013” In Revista Governança Corporativa**.** ([São Paulo]. v. 2, n.1 . pp. 94-115, abr. 2015) Disponível em: https://www.researchgate.net/publication/305056963_Caracteristicas_da_Producao_Academica_sobre_Governanca_Corporativa_no_Setor_Publico_Divulgadas_no_Web_of_Science_no_Periodo_1995-2013 . Acesso em: 02 mai. 2019.

Rieder, B. (2018) “Examinando uma técnica algorítmica: o classificador de bayes como uma leitura interessada da realidade” In Parágrafo ([S.l.], v. 6, n. 1, p. 123-142, jun. 2018. ISSN 2317-4919). Disponível em: http://revistaseletronicas.fiamfaam.br/index.php/recicofi/article/view/726 . Acesso em: 19 mar. 2019.

Rosa, C. S. (2018) “Estudo sobre as técnicas e métodos de análise de dados no contexto de Big Data” Disponível em: https://repositorio.ufu.br/bitstream/123456789/23710/1/EstudoTecnicasMetodos.pdf . Acesso em. 18 mai. 2019.

Sachet, M. e Silva, S. (2018) “Técnicas de Mineração de Dados Aplicadas na Análise de Dados de Fermentação de Vinhos”. Disponível em: https://repositorio.ucs.br/xmlui/handle/11338/3810 . Acesso em: 15 mar. 2019.

Sánchez, M. T. R., Alba-Ruiz, R. e Ramiro, T. (2014) “Bibliometric profile of RUSC. Universities and Knowledge Society Journal” In (Universities and Knowledge Society Journal) ([Granada], v. 11, n. 3, p. 42-62, jul 2014) Disponível em: https://www.researchgate.net/publication/264707137_Bibliometric_profile_of_RUSC_Universities_and_Knowledge_Society_Journal . Acesso: 12 mar . 2019.

Schubert, E. e Rousseeuw, P. (2018) “Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms”. Disponível em: https://arxiv.org/abs/1810.05691> . Acessoo em: 13 mar. 2019.

Sidhu, S. et al. (2014) “FP Growth Algorithm Implementation. International Journal of Computer Applications” In International Journal of Computer Applications**,** (maio. 2014) Disponível em: https://pdfs.semanticscholar.org/ef40/f70587ba1c653389c08cd21e96e387c5e6b8.pdf . Acesso em: 23 mai. 2019.