UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CURSO DE SISTEMAS DE INFORMAÇÃO

MARCO ANTONIO ROCHA BAUMANN

ÁRVORE DE DECISÃO PARA WEB ANALYTICS: UMA PROPOSTA DE DIRETRIZES PARA CLASSIFICAÇÃO DE MÉTRICAS DO GOOGLE ANALYTICS

FLORIANÓPOLIS
2017

Marco Antonio Rocha Baumann

ÁRVORE DE DECISÃO PARA WEB ANALYTICS: UMA PROPOSTA DE DIRETRIZES PARA CLASSIFICAÇÃO DE MÉTRICAS DO GOOGLE ANALYTICS

Trabalho de Conclusão de Curso submetido ao curso de Sistemas de Informação da Universidade Federal de Santa Catarina para a obtenção do Grau de Bacharel em Sistemas de Informação.

Área de Concentração: Business Intelligence
Linha de Pesquisa: Web Analytics

Orientador: Jorge Gustavo Sandoval Simão, Mestre

Coorientador: Roberto Carlos dos Santos Pacheco, Doutor

FLORIANÓPOLIS
2017

Marco Antonio Rocha Baumann

ÁRVORE DE DECISÃO PARA WEB ANALYTICS:
UMA PROPOSTA DE DIRETRIZES PARA CLASSIFICAÇÃO
DE MÉTRICAS DO GOOGLE ANALYTICS

Este TCC foi julgado adequado para obtenção do Título de "Bacharel em Sistemas de Informação", e aprovado em sua forma final pelo Curso de Sistemas de Informação da Universidade Federal de Santa Catarina.

Florianópolis, 05 de julho de 2017.

Frank Siqueira, Doutor
Coordenador do Curso

Jorge Gustavo Sandoval Simão, Mestre
Orientador
Universidade do Vale do Itajaí

Roberto Carlos dos Santos Pacheco, Doutor
Coorientador
Universidade Federal de Santa Catarina

Banca Examinadora:

José Leomar Todesco, Doutor
Membro
Universidade Federal de Santa Catarina

Denilson Sell, Doutor
Membro
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Agradeço primeiramente a minha família, meus pais e meu irmão, por todo o trabalho e luta para nos fornecer um futuro promissor, por meio de valores e oportunidades. Obrigado pai e mãe, Curt e Wanda, por serem meu Norte em tudo o que faço e por me apoiarem em meus projetos e devaneios, sem seu apoio muito do resultado seria impossível. Obrigado meu irmão, Victor, por todos estes anos de parceria e apoio.

Agradeço também aos outros “irmãos e irmãs” que estiveram presentes em minha vida e que também ajudaram a moldar quem sou hoje, amigos da PDP, Primatas, amigos de games espalhados pelo País e mundo, as madrugadas em claro nunca foram um desperdício de tempo.

Agradeço a minha namorada, Rackel, por ser paciente nos momentos que tinha que me dedicar 110% aos estudos e trabalho, e que agora sente na pele ao entrar na UFSC também.

Agradeço aos colegas de trabalho, em especial ao Victor, por ser mais do que parceiro de negócio, mas sim um grande amigo, aos colegas de faculdade, computação e sistemas que participaram desta minha jornada pela UFSC que se encerra depois de 10 anos.

Agradeço também a todos os professores envolvidos na minha vida acadêmica, por sua dedicação a profissão e importância para toda sociedade. Agradeço em especial para a professora Rose Linhares, que me introduziu a informática há muitos anos atrás, no ensino fundamental, o resultado do seu trabalho está presente aqui.

Por fim, agradeço a todas as pessoas que passaram em minha vida e contribuíram para que eu seja quem sou e alcançar meus objetivos.

RESUMO

Este trabalho de conclusão de curso centra-se na gestão do conhecimento para pequenas e médias organizações no que diz respeito a sua atuação na internet, elemento indispensável para a manutenção e melhoria da competitividade dos negócios digitais e seus websites. Dentro desse tema, aprofundamos nos conceitos de métricas coletadas por meio de web analytics – mais especificamente, com o uso do Google Analytics – e seu uso nas análises de performance de marketing digital através de Indicadores-chave de Desempenho (KPI) que desempenham papel fundamental quando tratamos avaliação de investimentos das organizações. É pretendido com este trabalho possibilitar a criação de uma árvore de decisão utilizando dados provenientes do monitoramento de acessos a um website, feito com Google Analytics. Para isto é sugerida uma diretriz passo-a-passo para coleta e transformação das métricas e dimensões coletadas pelo Google Analytics, sua classificação através do uso de técnicas de data mining, culminando na criação da árvore de decisão a ser utilizada pelos gestores de negócios. O processo sugerido busca utilizar ferramentas que não reflitam em investimento financeiro. Como resultado deste estudo espera-se incentivar o uso de métricas por gestores de pequenos e médios negócios digitais, fornecendo-lhes uma nova forma de avaliar as informações geradas pelo Analytics e fornecendo-lhes conhecimento que possibilite maior sucesso em tomadas de decisão.

Palavras-chave: Web Analytics. Métricas. Data Mining.

ABSTRACT

This final paper focuses on knowledge management for small and medium-sized organizations regarding their performance on the internet, an indispensable element for maintaining and improving the competitiveness of digital businesses and their websites. Within this theme, we delve deeper into the concepts of metrics collected through web analytics - specifically using Google Analytics - and its use in performance analysis of digital marketing through Key Performance Indicators (KPIs) that play a key role When we treat the organizations' investment assessment. It is intended with this work to enable the creation of a decision tree using data from the monitoring of access to a website, made with Google Analytics. For this, a step-by-step guide to collecting and transforming the metrics and dimensions collected by Google Analytics, its classification through the use of data mining techniques, is suggested, culminating in the creation of the decision tree to be used by the business managers. The suggested process seeks to use tools that do not reflect financial investment. As a result of this study it is hoped to encourage the use of metrics by small and medium digital business managers, providing them with a new way of evaluating the information generated by Analytics and providing them with knowledge that enables greater success in decision making.

Keywords: Web Analytics. Metrics. Data Mining.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

PDCA	Plan, Do, Check, Act
KPI	Key-Performance Indicator
BI	Business Intelligence
API	Application Programming Interface
GA	Google Analytics
ETL	Extract, Transform, Load
SGC	Sistema de Gestão do Conhecimento
OLAP	Online Analytical Processing
IP	Internet Protocol
MIT	Massachussetts Institute of Technology
ROI	Return of Investment – Retorno de Investimento
CPA	Custo por Aquisição
SEO	Search Engine Optimization
AITS	Administrative IT Systems and Services
NLP	Natural language processing – Processamento de Linguagem Natural
SVM	Support Vector Machine – Máquina de vetores de suporte
PDI	Pentaho Data Integration
ARFF	Attribute-Relation File Format
RAM	Random Access Memory – Memória de acesso aleatório
CART	Classification and regression tree
MLP	Multi Layer perceptron
PWPC	Probabilistic web page classifier
PCA	Principal Component Analysis
XML	eXtensible Markup Language
AVS	Attribute-value similarity measure – Medida de similaridade atributo-valor
AD	Árvore de Decisão

SUMÁRIO

1 INTRODUÇÃO

No mercado contemporâneo, a competitividade é uma das principais características para garantir a continuidade das organizações e seus negócios. Na da informação, a arma mais importante para a manutenção desta continuidade é, como o próprio nome diz, a informação: sobre o mercado, produtos, concorrentes, e a mais preciosa delas: sobre o cliente. É utilizando-se deste tipo de conhecimento que o negócio pode não só se manter, mas também se renovar, atendendo melhor a necessidade de seus clientes. Sendo assim imprescindíveis estes elementos na tomada de decisão de inovação que poderá resultar na fidelização seu público, evitando a perda deste para os seus concorrentes.

Seguindo o registro histórico realizado por Carvalho (2006) e sua tese de mestrado, a construção da internet no Brasil remonta aos anos 90, com a abertura do mercado brasileiro, durante o governo Collor, iniciou-se o processo de renovação tecnológica, e como consequência, a vinda da internet para o País após uma longa discussão sobre sua implantação no ambiente acadêmico brasileiro, concretizada pelo primeiro acesso acadêmico à internet no Brasil em fevereiro de 1991, seguido pela primeira versão da Rede Nacional de Pesquisa, projetada em 1992 e finalmente com a primeira conexão a longa distância estabelecida em 1993, entre São Paulo e Porto Alegre, e 1995 marcando o início das operações de provedores de internet.

Contudo, somente a partir de 2000, observa-se de fato a popularização da internet no Brasil, e com isto o surgimento de um novo mercado, os negócios digitais, também chamadas "pontocom", empresas que atuam parcial ou inteiramente através da internet, sejam eles *E-commerces*, *blogs* ou portais de notícia. Este novo meio de atuação, por mais que se diferencie em relação ao mercado convencional, é regido pelos mesmo princípio: a concorrência. Esta por sua vez coloca novamente em primeiro plano a competitividade, que pelo alto dinamismo e alcance do meio digital, se torna ainda mais determinante que no meio off-line.

Como resposta à esta nova realidade de mercado ocorre a popularização do uso de ferramentas de avaliação de desempenho de ações e métricas (FERNANDES; ROSA, 2013, p. 183), antes somente disponíveis para grandes empresas mediante altos investimentos. Empresas como *MicroStrategy*, Cognos, SAP e Oracle eram os grandes – e únicos – provedores deste tipo de conhecimento, oferecidos através de projetos que duram meses para serem implantados e com valores que podem custar dezenas de milhões de reais.

Atualmente se observa uma grande disponibilidade destes serviços para pequenas e médias empresas, alguns fornecidos de forma gratuita (FRIED; HANSSON, 2012). O grande

carro-chefe desta nova onda de web analytics é a ferramenta Analytics, da gigante Google, que tem concorrentes de peso, entre eles: WebTrends, SAS, QlikTech, MixPanel, Piwik, entre outras, que oferecem seus serviços gratuitamente, ou através de investimentos na casa de dezenas de dólares ao mês.

Tais ferramentas tornaram-se imprescindíveis para aferir e garantir a manutenção de negócios atuantes na internet. Porém torna-se necessária primeiramente a sua conceituação. Esta pode ser concisamente definida como:

> Mensuração, recolha, análise e geração de relatórios com dados da internet com o objetivo de compreender e otimizar a utilização de um *Website. Web Analytics* vai para além da simples medição de tráfego na internet, pois é usado em pesquisas de mercado e negócio, permitindo melhorar a eficácia e eficiência de um negócio online. Permite ajudar ainda as empresas a medir os resultados de determinadas campanhas de promoção e saber se a mesma está a correr como planejado. (LISBOA, 2012, grifo do autor).

A importância e a rápida expansão de utilização das métricas de tais aplicações se dá pela sua capacidade de geração de dados sobre o objeto analisado. Estes podendo ser utilizados como inputs para o ciclo *check*, do método iterativo de melhoria contínua PDCA – *Plan, Do, Check, Act* –, metodologia muito utilizada entre as empresas *pontocom*.

Os *outputs* dessas ferramentas possibilitam a tomada de decisão com base em dados estatísticos, de forma mais científica, em detrimento de tomadas de decisão puramente intuitivas. Estes *outputs* servirão de base para a concepção de novas oportunidades, ou até mesmo para mudanças de rotas na gestão estratégica dos negócios (LISBOA, 2012).

A presente pesquisa tem como foco o fornecimento de conteúdo relevante e necessário para a elaboração de um plano de monitoramento eficaz, evitando a ocorrência da sobrecarga cognitiva ou desinformação. Sua realização iniciará pelo levantamento das principais ferramentas de *web analytics* disponíveis no mercado, categorização de tipos de negócios digitais, avaliação das principais métricas e KPIs fornecidos pela ferramenta escolhida, catalogação dos dados gerados, suas classificações e relevância relativa ao contexto, resultando na criação de uma metodologia disponível aos gestores para que as situações de sobrecarga cognitiva sejam evitadas.

1.1 PROBLEMA DE PESQUISA

Com a utilização em grande escala dos softwares de web analytics como *Google Analytics*, ofertou-se aos gestores de negócios digitais informações sobre o desempenho de seus

websites e aplicativos. Porém estas são capazes de fornecer uma enorme quantidade de indicadores e informações diferentes sobre a o objeto monitorado, muitas vezes muito além do necessário aos gestores e/ou avaliadores de desempenho. Com este excesso de indicadores, os KPIs – *Key-Performance Indicators* – acaba por ocorrer o fenômeno contrário ao que se espera, como discorre Bittencourt (2013). De acordo com Andriotti (2008), apesar da abundância de informações, relata-se que o gestor tem a impressão de nunca as possuir em quantidade, ou qualidade, suficientes para a tomada de decisão. Pela conclusão de Andriotti evidencia-se a falta de conhecimento por parte dos gestores sobre a gerência de tais ferramentas e sobre quais informações se propõem a fornecer (FARRIS et al., 2013), que pode levar o gestor a sobrecarga cognitiva: que consiste na parcial falta de habilidade em processar de forma eficiente novas informações, principalmente, devido ao seu excesso (NAGASUNDARAM; DENNIS, 1993; GRISE; GALLUPE, 1999).

Outro conceito relacionado ao problema de pesquisa é o conceito da desinformação, o qual Pinheiro e Brito (2014) definem por informações que não agregam valor ou conhecimento, e que são disponibilizadas em fluxo ininterrupto, podendo aturdir e diluir a capacidade de processamento de um alvo em questão. Esta situação tem como consequência a redução da eficiência de ações de organizações no ambiente digital, podendo culminar no comprometimento do desempenho e dos objetivos desejados pela organização. Por fim, tem-se como principal motivação para realização deste trabalho a frase escrita por Avinash Kaushik:

> Existe uma profunda falta de conhecimento prático real no mercado. Mais importante ainda, existe uma falta de pessoas e práticas que permitam aos negócios digitais obter conhecimentos que resultem em ações as quais produzam diferenciação estratégica entre eles e seus concorrentes. (KAUSHIK, 2007, p. 7, tradução nossa).

1.1.1 Solução proposta

Como solução para o problema apresentado, o objetivo de pesquisa é a proposta de uma diretriz para a construção de uma árvore de decisão utilizando softwares gratuitos, tendo como base dados fornecidos pela ferramenta de *web analytics* Google Analytics em sua versão gratuita, possibilitando aos gestores de negócios digitais uma alternativa mais simples para a aquisição de conhecimento sobre seus negócios digitais por meio de BI e *data mining* sem a necessidade de investimentos financeiros e consequentemente tornando conhecimento mais acessível ao mercado e às pequenas e médias empresas.

1.1.2 Delimitação de escopo

Será utilizado nesta pesquisa o software de *web analytics Google Analytics* em sua versão gratuita, devido a abrangência de seu uso, fato verificado durante a pesquisa deste trabalho. A forma utilizada para a coleta dos dados e informações será a API – *Application Programming Interface* – pública do *Google Analytics*, respeitando as políticas de uso estabelecidas pela fabricante do software. Para a manipulação dos dados coletados e finalmente classificação destes são utilizadas ferramentas gratuitas e conhecidas no meio acadêmico, Pentaho e Weka, respectivamente. Este trabalho limita-se a avaliar a viabilidade de criação da árvore de conhecimento, por meio das ferramentas acima descritas, deixando a avaliação da qualidade do conhecimento gerado para trabalhos futuros.

1.2 JUSTIFICATIVA

Este trabalho objetiva a utilização de técnicas de data mining sobre os dados e informações ofertados pela ferramenta *Google Analytics*, uma das mais utilizadas no mercado mundial (DATANYZE, 2017), facilitando aos gestores de negócios digitais a avaliação destes dados para realizar as mais variadas decisões sobre seus negócios. Esta facilitação ocorre, pois, na aplicação do algoritmo de classificação ocorre a seleção das variáveis – as métricas e KPIS – que tem maior ganho de informação, ou seja, têm maior relação com os objetivos do website avaliado. Com isto há a diminuição do universo de métricas a serem avaliadas para apenas o conjunto reduzido e de alta relevância, simplificando o processo e poupando tempo e esforços dos gestores para entender melhor se comporta o cliente ideal de seu website.

1.3 OBJETIVOS

1.3.1 Objetivo geral

O objetivo deste trabalho é a proposta de um conjunto de diretrizes eficazes e que não impliquem em investimentos financeiros para elaboração de uma árvore de decisão utilizando dados de visitação coletados pelo Google Analytics. Esta árvore poderá ser utilizada por gestores de pequenas e médias organizações na tomada de decisão de investimentos em seus negócios digitais.

1.3.2 Objetivos específicos

- Conceituar as principais ferramentas de web analytics por suas principais características, justificando a escolha do *Google Analytics* como ferramenta a ser utilizada;
- Conceituar e classificar dos métricas e *KPI*s;
- Criar um processo eficaz para a classificação das métricas coletadas de seu website utilizando softwares gratuitos;
- Gerar uma árvore de decisão do website/aplicação avaliada.

2 METODOLOGIA

2.1 METODOLOGIA DA PESQUISA

Seguindo os preceitos de Wazlawick (2008), para a obtenção dos objetivos, tanto gerais quanto específicos, deste trabalho foram realizadas pesquisas bibliográficas sobre os temas abordados, desde os fundamentos da gestão do conhecimento, conceitos de dados, informação e conhecimento, sobre a natureza de *web analytics* modernos, tipologia de websites e suas principais métricas e indicadores de desempenho.

Sob a ótica de sua natureza, este trabalho tem com algo a geração de conhecimento para a aplicação prática de solução para um problema específico, classificando-se assim o trabalho como uma pesquisa aplicada.

Sob o ponto de vista de seus objetivos, esta pesquisa é exploratória, pois envolver um levantamento bibliográfico, análise de conceitos e classificações de autores e validação da solução proposta.

2.1.1 Procedimentos metodológicos

O trabalho será realizado, iniciando-se com uma revisão bibliográfica sobre o tema de métricas de monitoramento digital. Com objetivo de se obter o que há de mais recente sobre os conceitos referidos neste trabalho, tornar-se-á necessária pesquisa em publicações não somente em português, mas em outras línguas, a pesquisa poderá abranger conteúdo e materiais - apresentações e publicações digitais - elaborados por organizações e pessoas proeminentes nas áreas de métricas, *KPIs* e *web analytics*.

Elaborada a conceituação, inicia-se a pesquisa de mercado buscando definir quais as principais ferramentas de coletas de dados utilizadas e justificando a escolha do *Google Analytics* como ferramenta utilizada.

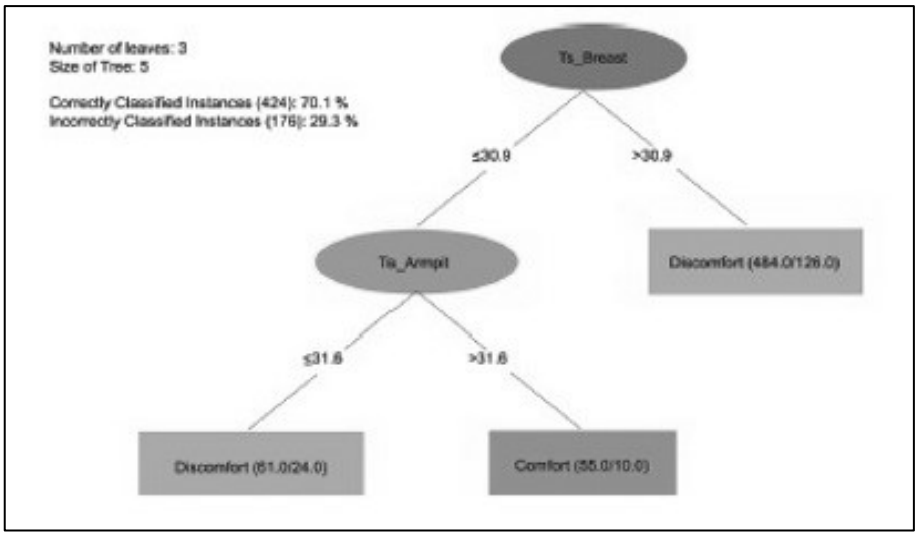
Como finalização da fundamentação teórica são conceituadas as métricas e *KPIs* analisados pelo GA, aprofundando-se em suas diferenças e, segundo autores, as principais características que as definem, chegando às suas classificações e relacionando-as a grupos de contexto.

Na etapa de desenvolvimento é elaborada a tarefa de classificação de data mining destas métricas coletadas pelos *web analytics* para criação da árvore de decisão para ocorrência de um evento-objetivo do website, também chamado de conversão. O processo inicia-se pela coleta dos dados advindos do *Google Analytics* através de sua API, utilizando para este fim o software

Pentaho *Data Integration*, ferramenta de ETL, Analytics e Big Data em sua edição comunitária, realizando a manipulação dos dados para que se adequem à especificações necessárias do software Weka, ferramenta open-source oferecida pela universidade de Waikato, Nova Zelândia, para aplicação de análise de conhecimento e data mining, onde é realizada a classificação dos dados utilizando o algoritmo de classificação C4.5 em sua implementação J48, presente no Weka. Após estes passos é gerada a arvore de classificação das métricas providas pelo Google Analytics indicando sua relevância na ocorrência de um evento almejado.

O resultado esperado é a verificação da efetividade das diretrizes para criação de um artefato de conhecimento de fácil entendimento que poderá ser utilizado por gestores não-técnicos em tomadas de decisão quanto a futuros investimentos em seus negócios digitais. A árvore de decisão terá sua estrutura semelhante à apresentada na Figura 1.

Figura 1 - Exemplo de árvore de decisão



Fonte: Maia et al. (2013).

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados conceitos relevantes relacionados aos elementos, a motivação do uso e os procedimentos envolvidos na avaliação de desempenho de websites de negócios digitais e nas formas de sua melhoria, sendo feita uma revisão bibliográfica para consequente definição do escopo da solução proposta.

3.1 DADO, INFORMAÇÃO E CONHECIMENTO

Segundo diversos autores, dados e informações são considerados sinônimos, porém para melhor andamento da presente pesquisa, devemos seguir a conceituação de autores que os distinguem, e assim tornando mais clara a compreensão dos elementos. Dados e informação são conceitos profundamente ligados ao objeto de estudo desta pesquisa, por isto a necessidade de esclarecer a distinção entre eles e suas representações análogas na mensuração de desempenhos de websites.

Inicia-se a conceituação de dado pelas palavras de Houaiss (2001, p. 903), que se segue:

> Existe uma profunda falta de conhecimento prático real no mercado. Mais importante ainda, existe uma falta de pessoas e práticas que permitam aos negócios digitais obter conhecimentos que resultem em ações as quais produzam diferenciação estratégica entre eles e seus concorrentes. (KAUSHIK, 2007, p. 7, tradução nossa).

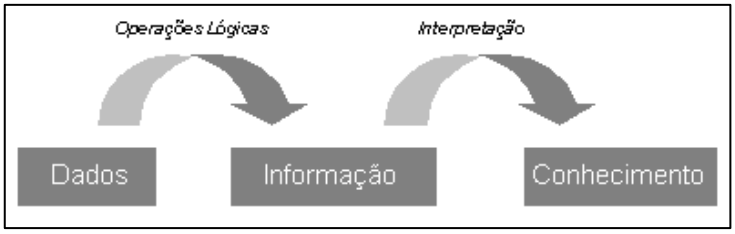
Outro autor, Santos (2009), descreve dados como elementos base para a criação de informação, sendo este tudo o que é captado por um sensor. Contudo, a definição que mais se relaciona com o contexto desta pesquisa é descrita por Rabaca e Barbosa (1995), onde dado é definido como “elemento numérico, conhecido ou obtido por método de coleta apropriado, que serve de base para processo de análise”.

O que é visto como consenso entre os autores, é que dados são considerados as unidades básicas para qualquer sistema de conhecimento, e por este valor semântico básico, carecem de significado próprio, quando avaliados de forma separada. Por sua vez, sendo entendidos como peças agregáveis, quando aplicados operações lógicas em conjunto a outros dados, há a possibilidade de associação de seus valores e com isto a criação de significado aos dados referentes ao objeto mensurado em questão.

Quando a ocorrência de significado sobre um conjunto de elementos base, ou numéricos, define-se o conceito de informação, que segundo Serra (2007) informação é a resultante do processamento, manipulação e organização de dados, de tal forma que represente

uma modificação (quantitativa ou qualitativa) no conhecimento do sistema (humano, animal ou máquina) que a recebe. Ou seja, um elemento que carrega valor semântico às pessoas ou sistemas que venham a interpretá-lo com possibilidade de modificar ou gerar um terceiro elemento: o conhecimento. Esta relação entre estes três elementos é evidenciada por meio do exposto na Figura 2.

Figura 2 - Representação de dados, informação e conhecimento



Fonte: adaptado de disponível em: <http://fdr-sig-jonathanroger.blogspot.com.br/2009/11/dados-e-informacoes.html>.

Como terceiro elemento desta linha temos o conhecimento. O esforço de definição do conhecimento é uma constante na sociedade acadêmica que, por reflexo, resulta em inúmeras definições diferentes, iniciando-se pela definição de Platão, onde o conhecimento consiste em uma crença verdadeira e justificada de um sujeito, também chamado de cognoscente, em relação a um objeto, o cognoscível. Ou seja, a elaboração é diretamente dependente do ator envolvido no processo. Há autores, porém, que discordam desta definição, entre eles, Gettier, que questionam a suficiência das três características como condição para a existência de conhecimento.

Autores atuais ampliam o conceito de conhecimento para novas formas, sempre visando a importância do sujeito atuante no processo, como Sveiby (1998), relaciona o conhecimento à capacidade de ação. Porém, a definição mais própria para este trabalho vem de Nonaka e Takeuchi (1997), que declaram que "o processo de criação de conhecimento diz respeito a crenças e compromissos" e sempre está ligado a ações, atitudes e a intenção específica do cognoscente, "um processo humano dinâmico de justificar a crença pessoal em relação 'a verdade'".

3.2 NEGÓCIOS DIGITAIS – DOT.COM

Negócios digitais, ou também chamados de empresas *dot-com*, são tipos de organizações que realizam seus negócios de forma majoritária ou completamente através da internet (BEYNON-DAVIES, 2004), onde há compartilhamento de informação pela internet,

ou seja, são empresas ou organizações que dependem fundamentalmente de seus *websites* ou aplicativos baseados na internet para realização de suas operações de negócios e interação com sua clientela, sendo o exemplo mais evidente as lojas online, também chamados de e-commerce.

Por utilizar a internet como plataforma de oferta de serviços e produtos, por um lado possibilitando acesso a maiores mercados do que no meio "offline", por outro há a elevada concorrência que a internet trás. Como consequência disto, cria-se uma maior necessidade que a organização avalie constantemente a forma com que mercado se comporta e busque formas de avaliar o seu desempenho e de seus concorrentes através do BI com objetivo de manter e melhorar a sua competitividade e assegurar a sobrevivência da organização.

3.3 BUSINESS INTELLIGENCE

Mesmo havendo diferentes definições dadas por diferentes autores, desde centradas puramente nos processos e instrumental envolvidos, como definido por Berson e Smith (2002), BI inclui diversos softwares para Extração, Transformação e Carregamento, *data warehousing*, busca em bases de dados e relatórios, OLAP, análise de dados, *data mining* e visualização, chegando a conceituações diretamente ligadas aos resultados, como o resultado de análise profunda de dados de negócios detalhados, incluindo tecnologias de banco de dados e aplicações, junto a processos de analíticos (GANGADHARAN, 2004) e Zeng et al. (2006), sendo BI "O processo de coleta, tratamento e difusão de informação que tenha um objetivo, a redução da incerteza na realização de decisões estratégicas".

> *Business Intelligence* (BI) é definida como um conjunto integrado de ferramentas que dão suporte a transformação de dados em informações de forma a subsidiar a tomada de decisão. Entretanto, organizações tem dependem do uso mais abrangente do BI que compreende também a habilidade de analisar informações sob o contexto de necessidades particulares e o uso de tecnologias de gestão do conhecimento de forma a acelerar o processo de criação de conhecimento para decisão. (SELL et al., 2012).

Todas estas definições orbitam o conceito de que o *business intelligence* é a habilidade de uma organização converter processos, ações e interações em conhecimento, e garantir a entrega deste conhecimento às pessoas certas, da forma correta, no momento correto. Como consequência direta do BI temos a possibilidade de desenvolvimento de novas oportunidades e perspectivas para a organização (KUMARI, 2013). Junto a este processo estão práticas e ferramentas de ETL e data warehouse, sendo o papel deste último no contexto deste trabalho desempenhado pelo Google Analytics que, mesmo que não apresente todas as qualidades e

processos de uma aplicação de data warehouse, atende ao conceito definido por Zeng como uma ferramenta de BI para o fornecimento de informação sobre um website ou aplicação web.

Power (2007) já utiliza a definição de Dresner de Business Intelligence como conceitos e métodos para melhoria da tomada de decisão de uma organização por meio de sistemas baseados em fatos. Tais sistemas baseados em fatos são as ferramentas de data warehousing ou datamarts, como descritos por autores supracitados. O BI cobre estas ferramentas e outras técnicas, como mineração de processos, benchmarking, data mining.

O BI por si pode ser aplicado a inúmeras situações, sendo este termo, como escrito por Kobielus (2010), business apenas a forma mais generalizada dos dados analíticos entregues a usuários por meio de relatórios e *dashboards*. Abaixo dele se encontram outras inteligências como de mercado, competitiva, social. O autor ainda conclui que, em sua visão, não há diferenciação prática entre inteligência e analítica, sendo possível substituir o termo por outro em quaisquer das definições sem alteração de significado. O autor também descreve o que pode ser considerada a evolução das técnicas de BI, partindo inicialmente de um ambiente onde os dados e informações eram em sua grande maioria estruturadas, e hoje trabalha-se com dados semi ou não-estruturados, principalmente no âmbito das redes sociais e outros que envolvam principalmente interação interpessoal.

Estes dados e informações utilizadas pelo BI podem ser agrupados em três grandes grupos, não-estruturados, semiestruturados e estruturados:

> **Não-estruturados:** são informações que não tem um modelo pré-definido de organização e/ou não são organizadas em uma ordem pré-estabelecida, fazendo com que este tipo de informação seja armazenado entre os dados com esta característica temos e-mails, transcrições de conversas telefônicas, tabelas de dados, documentos. Grimes (2008) cita em seu artigo que aproximadamente 80% a 85% das informações de empresas úteis para BI sejam deste tipo.

> **Semiestruturados:** são informações que mesmo não tenham uma estrutura definida formalmente, porém já contém meta dados e elementos semânticos em seu corpo, tornando-as conhecidas como estruturas auto descritas (UNIVERSIDADE DE CHICAGO, 2017). Não sendo, desta forma, possível determinar que todas as entidades de mesmo tipo, contenham as mesmas quantidades de atributos. Exemplos de dados semiestruturados mais comuns são o XML e o JSON.

Estruturados: São os tipos de dados comumente utilizados em aplicações computacionais, com modelo formalmente definido, podendo ser representações alfanuméricas, normalmente armazenadas em bancos de dados relacionais. Este é o tipo de dado armazenado pela ferramenta Google Analytics, portanto utilizado para a elaboração da árvore de decisão.

Dentre os instrumentais envolvidos no Business Intelligence, destaca-se o uso do data mining para a geração da árvore de decisão neste trabalho, sendo importante sua conceituação e de suas tarefas.

3.4 DATA MINING

Data mining, ou mineração de dados, é o nome dado ao processo computacional que tem como objetivo a descoberta de conhecimento em volumes de dados e informações envolvendo métodos de inteligência artificial, estatísticas e banco de dados, como descrito por Clifton (2015). Este processo tem como objetivo extrair informações de uma estrutura de dados e transformá-la em uma estrutura compreensível os agentes que irão posteriormente utilizá-las (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A mineração de dados pode detectar padrões como grupos de dados com mesmas características (análise de cluster), detecção de anomalias, e dependências (mineração por regras e associação), podendo estes serem utilizados em análises posteriores e para análise preditiva.

O uso do *data mining* há muitas décadas, onde métodos antigos de identificação de padrões, entre eles o teorema de Bayes e a análise de regressão, isto em meados dos séculos XVI e XVII. Com os avanços exponenciais nos setores tecnológicos e do poder de processamento, os grupos de informações aumentaram em complexidade e tamanho, com isto, a manipulação destas informações também foram melhoradas por meio do processamento de dados indireto e automatizado, auxiliadas com descobertas no setor de computação.

Suas aplicações em tempos atuais são nas mais diversas áreas, entre elas negócios, jogos e minerações de padrões, entre outras, a seguir estão listadas utilizações do data mining nas áreas:

Negócios: Utilizada para avaliar históricos de transações com objetivo de buscar características e tendências nos dados, são usados algoritmos de reconhecimento de padrões em grandes quantidades de informações para auxiliar na descoberta de

conhecimento estratégicos dos negócios avaliados (O'BRIEN; MARAKAS, 2011). Como exemplo de seu uso, para a indústria da propaganda o data mining é uma ferramenta muito útil para catalogar informações provenientes do mercado, podendo identificar padrões de consumo de clientes e identificar os mais propensos a responder positivamente uma campanha ou propaganda por e-mail (BATTITI; BRUNATO, 2011)

Jogos: No xadrez, com objetivo de extrair as estratégias utilizadas por jogadores humanos contra máquinas, foram utilizados experimentos com sistema de análise pré-calculada da posição das peças do tabuleiro, que combinado a um estudo intensivo do conhecimento adquirido, gerou uma sequência de padrões preditivos para alimentar a inteligência de um jogador não-humano (O'BRIEN; MARAKAS, 2011).

Mineração de Padrões: Técnica da descoberta de conhecimento que envolve a descoberta de padrões em um determinado universo de dados, ou seja, associações entre seus registros. Um uso prático da mineração de padrões, segundo a *National Research Council* (2008), é a identificação de atividades terroristas, mesmo que estas atividades tenham sejam pouco perceptíveis no universo de dados.

3.4.1 Tarefas do data mining

A mineração de dados, como processo, é definida por Groth (1998) é aquele responsável pela descoberta automática de conhecimento. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) a mineração de dados possui seis tipos comuns de tarefas, brevemente descritas na lista a seguir:

- **Detecção de Anomalias:** identificação de registros não-usuais, que podem ser interessantes ou erros de informação que requerem algum tipo de investigação.
- **Aprendizado por associação (Modelo de Dependência):** são procurados relacionamentos entre variáveis, como por exemplo um supermercado que pode coletar informações sobre os hábitos de compra de um cliente.
- **Clusterização:** a tarefa de descobrir grupos e características do grupo de informações-alvo, que são de uma forma ou outras semelhantes, sem utilizar as estruturas conhecidas do grupo de informações em questão.

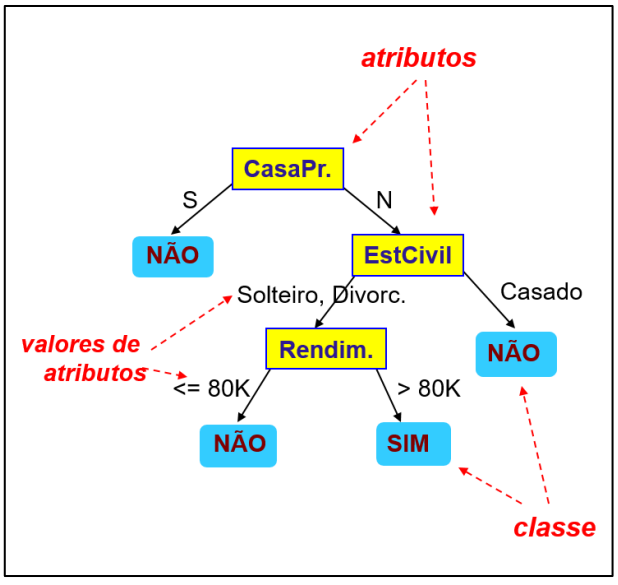
- **Classificação:** a tarefa de catalogar a informações em categorias conhecidas, tal como um gerenciador de e-mails que cataloga novos e-mails como legítimos ou spams.
- **Regressão:** busca encontrar uma função que modela a informação baseado no que se aprender com o último erro.
- **Sumarização:** Provê uma representação mais compacta do grupo de informações, incluindo visualização e geração de relatórios.

3.5 ÁRVORES DE DECISÃO

Árvores de decisão são ferramentas de suporte a decisão, criada a partir de um conjunto de dados, comumente chamado de conjunto de treinamento. Sendo um grafo em forma de árvore, um dos métodos mais amplamente usados e práticos para inferência indutiva sobre um conjunto de dados, suas relações são representadas por um conjunto de regras "se-então" para facilitar a legibilidade humana (MITCHELL, 1997). Sua representação se dá por meio de três elementos-base:

- **Nodos:** representam atributos da entrada;
- **Arcos:** correspondem ao valor de um atributo;
- **Nodos-folha:** provê a classificação da instância, ou entrada no conjunto de dados.

Figura 3 - Exemplo de árvore de decisão



Fonte: Bogorny (2015).

A árvore de decisão é construída por meio da aplicação de algoritmos de aprendizados sobre um conjunto de treinamento, sendo dois dos mais conhecidos o algoritmo ID3 (QUINLAM, 1986) e sua posterior extensão C4.5, este último utilizado na solução proposta deste trabalho. Bogorny descreve em seu material os passos realizados para a construção da árvore, que são:

1) Seleciona-se um atributo como sendo nodo raiz;
2) Arcos são criados para todos os diferentes valores do atributo selecionado no passo 1;
3) Se todos os exemplos de treinamento (registros) sobre uma folha pertencerem a uma mesma classe, esta folha recebe o nome da classe. Se todas as folhas possuem uma classe, o algoritmo termina;
4) Senão, o nodo é determinado com um atributo que não ocorra no trajeto da raiz, e arcos são criados para todos os valores. O algoritmo retorna ao passo 3.

Para a escolha dos atributos que serão utilizados como nodo na árvore e sua posição (mais próxima da raiz ou das folhas) é baseada na Teoria de Informação de Shannon, mais especificamente nos conceitos de Entropia e Ganho de Informação.

Entropia: Definida como a quantidade necessária de informação para identificar a classe de um caso. Dada pela equação:

$$Entropia(S) = -(p_1 \log_2 p_1 + p_2 \log_2 p_2 + \ldots + p_n \log_2 p_n)$$

Onde:
S é o conjunto de amostras (registros);
n é o número de valores possíveis da classe;
$\mathbf{p_i}$ é a proporção de amostras da classe *i* em relação ao total de amostras.

Ganho de Informação: É a redução esperada da entropia ao utilizarmos um atributo na árvore. Dada pela equação:

$$Ganho(S, A) = Entropia(S) - \sum((|Sv| / |S|) * Entropia(Sv))$$

Onde:
Ganho (S, A) é o ganho do atributo A sobre o conjunto S;

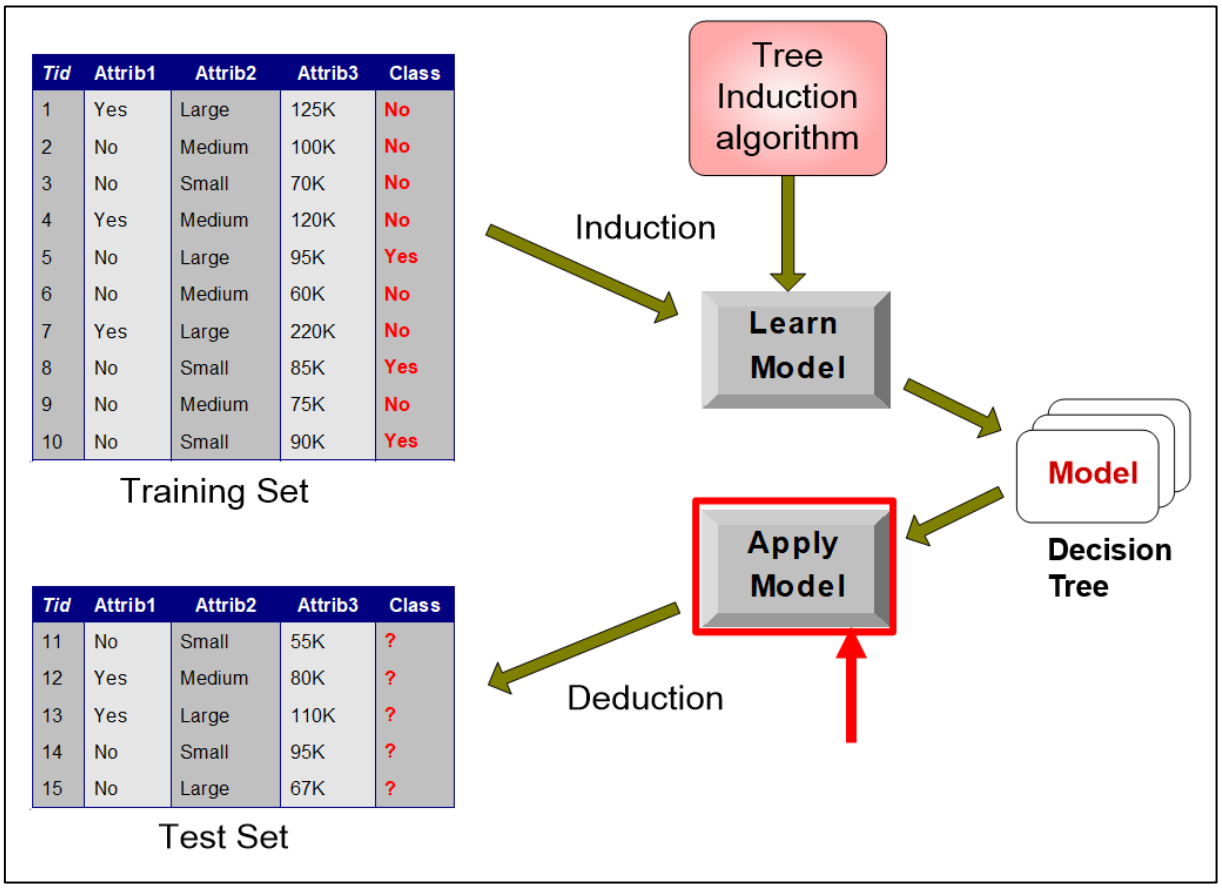
Sv é o subconjunto de S para um valor do atributo A;

|Sv| é o número de elementos de Sv;

|S| é o número de elementos de S.

Criada a árvore, esta pode ser utilizada para classificar um conjunto de entradas de dados não classificada previamente, chamado de conjunto teste, sendo atribuída a esta classificação uma possibilidade de erro, dependendo das características do conjunto de treinamento e configuração do algoritmo de aprendizado utilizado para criação da árvore.

Figura 4 - Visão geral de classificação utilizando árvore de decisão



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Fonte: Bogorny (2015).

Segundo Mitchell (1997), os algoritmos ID3 e C4.5 têm melhores resultados na elaboração de árvores de decisão de tamanhos considerados pequenas, justificando o fato com base no paradigma da lâmina de Occam, “prefira a hipóteses mais simples que explica os dados”, aplicando na árvore de decisão a poda de nodos com baixo ganho de informação, reduzindo assim sua complexidade e mantendo a legibilidade do artefato.

3.6 BENCHMARKING E O PLANEJAMENTO ESTRATÉGICO

Organizações inovadoras estão sempre buscando formas de melhorar seus produtos e otimizar seus processos, porém para ter de fato a constatação de que há aumento na qualidade ou na eficiência, devem ser estipulados parâmetros e é a na comparação dos indicadores do negócio com tais parâmetros que reside o benchmarking. Como descrito por Carlini e Vital (2004), o benchmarking auxilia empresas a identificar seus pontos fortes e fracos, para então desenvolver – e melhorar – suas estratégias de inserção e permanência em um mercado.

Em seu artigo, o autor estrutura o *Benchmarking* em 3 etapas: seu planejamento, sua execução e a implantação de melhorias, culminando na maximização da competitividade organizacional. As ferramentas de *web analytic* tem seu principal momento de atuação na execução, auxiliando na coleta de dados sobre o negócio digital.

No entanto, a proposta trazida por este trabalho foca em auxiliar a elaboração da primeira parte do benchmarking, onde são determinados os indicadores a serem avaliados e comparados por meio da ordenação dos indicadores de acordo com seu grau de relevância e correlação com os objetivos – definidos como conversões – do negócio digital avaliado. Com isto conseguimos determinar quais são os principais indicadores de desempenho, podendo diminuir o número de variáveis avaliadas, simplificando o processo.

3.7 IMPORTÂNCIA DO WEB ANALYTICS PARA NEGÓCIOS DIGITAIS

Pelo fato de um negócio digital ter como principal meio de atuação a internet, a prática de *web analytics* acaba por atuar como principal meio de mensuração de desempenho da organização, coletando informações quanto a sua interação com o mercado, seus clientes e possíveis novos clientes e gerando medidas de desempenho do negócio. Estas, como descritas por Hronec (1994), são os maiores indicadores da saúde da organização, quantificando e qualificando ações da organização que visam o atingimento de seus objetivos estratégicos.

A utilização desta ferramenta para mensurar o desempenho pode trazer benefícios a gestão do negócio, fornecendo um conjunto de dados reais sobre a satisfação de seus clientes, resultados de ações aplicadas à estrutura digital da organização e retorno sobre investimentos realizados em marketing digital, ou seja, tornando-se a principal fonte de informação para saber quão bem a organização está cumprindo seu objetivo e para direção estratégica do negócio digital.

3.8 FERRAMENTAS DE WEB ANALYTICS

3.8.1 Panorama histórico

Segundo Kaushik (2007), as ferramentas de web analytics têm como motivação de existência a necessidade de aferimento dos erros ocorridos nos servidores na *World Wide Web*, a fim de descobrir se o sistema está funcionando corretamente. Posteriormente descobriu-se a possibilidade de recolher informações mais detalhadas sobre as solicitações recebidas por estes servidores, dados como o endereço de IP do requerente, identidade de seu navegador, sistema operacional, hora da requisição, entre outros dados. Com isto este aferimento tornou-se foco de interesse do público não-técnico, termo utilizado pelo autor para profissionais não diretamente ligados à tecnologia. Kaushik tem como data de criação de tais ferramentas o ano de 1995, onde Dr. Stephen Turner, na época participante do laboratório de estatística da universidade de Cambridge, publicou o software *Analog* em sua versão 0.9b que foi considerado o primeiro programa de análise de registro, o pai das ferramentas de web analytics modernas.

No ano 2000 surgiram novos players no setor de web analytics, desta vez com foco comercial, entre elas Accrue, WebTrends e Coremetrics, as quais estenderam as funcionalidades originais do programa *Analog*, fornecendo a seus usuários gráficos e outras ferramentas para avaliação dos dados. Tais ferramentas traziam maiores possibilidades aos seus usuários, mas com o revés de serem ferramentas pagas, o que limitou a disseminação do uso de *web analytics* na internet, que na época consistia em aproximadamente 17 milhões de websites, segundo estudo realizado conjuntamente pelo MIT, Hobbes Internet Timeline e Pingdom (INTERNET LIVE STATS, 2016).

Kaushik (2007) marca o ano de 2005 com uma grande mudança no panorama de web analytics com a aquisição da empresa Urchin pela Google e a posterior lançamento de sua própria ferramenta de análise *Google Analytics* de utilização gratuita, ocorrido em 2006. Desde então, as mais diversas inovações têm sido implementadas aos softwares analíticos com objetivo de munir os gestores com informações sobre seus negócios digitais.

3.8.2 Desafio atual

Com a entrada do *Google Analytics*, o mesmo simplesmente explodiu, pois agora qualquer um que quiser dados sobre seu website pode os ter gratuitamente (KAUSHIK, 2007), porém esta disponibilidade de nada vale se o gestor ou avaliador não se utilizar de um plano ou método para avaliar as métricas de sucesso relevantes ao seu tipo de negócio, atribuir confiança

aos *ROIs* gerados por tais ferramentas. Por fim, evitar o que o próprio autor chama de "Paralisia por análise" (KAUSHIK, 2007, tradução nossa).

Uma solução amplamente praticada no mercado atual é a predefinição e padronização de *KPIs*, onde o mesmo conjunto de indicadores de desempenho são designados para qualquer tipo de negócio digital. Esta aproximação pode solucionar de maneira temporária o problema, em um cenário onde há pouco ou nenhum conhecimento sobre este tipo de avaliação, mas com a evolução dos trabalhos de pesquisa e análise, fica desvelado a necessidade de adaptação destes *KPIs* para a realidade do negócio, adaptando-os para as particularidades e realidade do mercado onde o negócio está inserido.

3.9 CLASSIFICAÇÃO DE FERRAMENTAS DE WEB ANALYTICS

Partindo de um simples *software* de análise de registros de servidores, como descrito por Kaushik (2007), ferramentas de *web analytics* evoluíram e se diferenciaram no que diz respeito a metodologia usada para o desenvolvimento da ferramenta. Ribeiro et al. (2012, p. 21) as classifica em seu *e-book* em dois grandes grupos, determinando o foco de coleta e avaliação dos dados de cada um e também listando exemplos de softwares disponíveis no mercado com tais características. Estes dois grupos são:

> **User Centric – Análise centrada no usuário:** como o nome propriamente diz, o objeto de estudo é o usuário – ou visitante – do *website*, com objetivo de determinar e avaliar hábitos, costumes e tendências relacionadas ao perfil do usuário ou ao perfil ou segmento de mercado do qual ele faz parte, não se restringindo a coletar dados somente pelas visitas realizadas a um website em específico. Exemplos de software listados pelo autor, os quais se utilizam desta metodologia são Nielsen Online e ComScore.

> **Website Centric – Análise centrada no website:** baseia-se em senso para fornecer informações detalhadas sobre o *website* avaliado, armazenando informações sobre cada uma das visitas realizadas ao objeto analisado, gerando dados de alta confiabilidade. Infelizmente há a restrição de apresentação de dados e informações somente do *website* objeto da análise, o que impossibilita a visão do mercado e comparação de desempenho com outros *websites*. Felizmente, como descrito por Google (2017), já está sendo disponibilizada na ferramenta *Google Analyti*cs a funcionalidade de avaliação comparativa com participantes do mercado.

Seguindo Mehta (2015), que em seu artigo classifica web analytics em 4 grandes grupos sob a perspectiva do foco de análise, sendo que ferramentas disponíveis no mercado podem abranger um ou mais dos grupos definidos pelo autor. Estes grupos são classificados em:

Análise de Usuário: centra-se na aquisição de dados sobre o usuário que está realizando a visita, obtendo informações como sua idade, localização, preferências e qualquer outra informação que possa ajudar na determinação do perfil deste visitante, além destes pontos também se entende como análise de usuário o monitoramento de como ele visita – dispositivo usado, por exemplo – seu *website*, que informações absorveu e dificuldades que possa ter tido durante a visita.

Análise de Fonte de Tráfego: esta classe de análise foca em descrever o fluxo de entrada de visitantes, buscando evidenciar como as pessoas chegaram ao website, sendo este tráfego comumente dividido em 3 grupos: tráfego de mecanismo de busca, tráfego direto e tráfego de referência, descritos no Quadro 1:

Quadro 1 - Categorias de fontes de acesso

Tráfego de mecanismo de busca	Inclusas neste grupo estão visitas provenientes de websites indexadores, como Google, Yahoo e Bing, onde o usuário realiza uma pesquisa sobre algum assunto e acessa os websites listados como relacionados ao assunto pesquisado. Podem ser especificados como pago, onde a visita resulta em pagamento ao website indexador, e orgânico, onde isto não ocorre.
Tráfego direto	Ocorre quando o visitante digita o endereço do website visitado diretamente no navegador, sem o direcionamento de nenhum website indexador ou de terceiros.
Tráfego de referência	Este tráfego é proveniente de website que contenham um direcionamento ou referência ao website avaliado, sendo que o website-fonte não é um indexador de conteúdo. Dentro deste grupo estão blogs, redes sociais.

Fonte: Google (2017).

3.9.1 Análise comportamental

Busca compreender o fluxo de navegação do visitante e suas ações durante a visita, com objetivo de monitorar e gerar dados sobre a visita que possam resultar em melhoria da experiência dos usuários futuramente, descrevendo tendências e padrões de ações desejadas ou indesejadas de seus usuários e buscando entender como este usuário interage com as informações e conteúdos oferecidos a ele.

3.9.2 Análise de aquisição

Objetiva analisar e otimizar o investimento em tráfego ao *website*, envolvendo todas as ações possível para tal, desde tráfego de mecanismo de busca pago, redes publicitárias, e-mail marketing, redes sociais e outras formas de atração de público, ajudando a criar uma relação entre o investimento realizado e o retorno gerado ao website em questão.

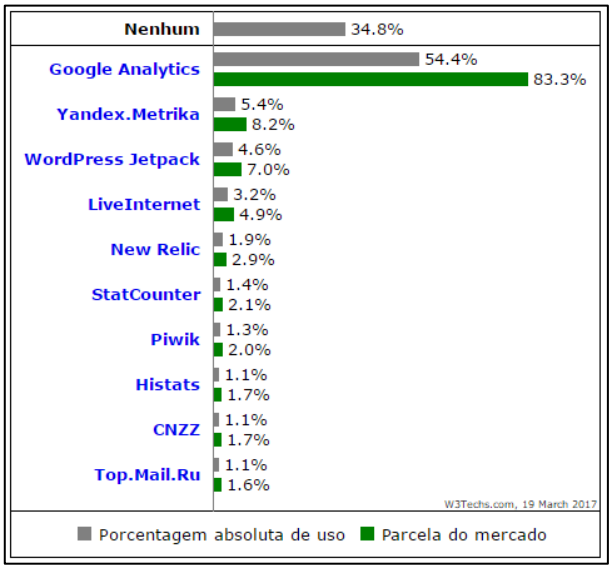
3.10 GOOGLE ANALYTICS

Para o desenvolvimento deste trabalho, utilizaremos a ferramenta *Google Analytics*, de propriedade da Google Inc., gigante do mercado de tecnologia e internet, para realizarmos os estudos, detalhando e classificando as métricas e indicadores-chave de performance por ela gerados.

Esta é uma decisão tomada com base em três motivos: o fato de ser uma ferramenta gratuita; abranger os quatro tipos de análise de *web analytics* (GOOGLE INC., 2017); responder pela maior parcela do mercado de ferramentas de web analytics, conforme demonstrado pelos estudos da W3Techs (2017) e Datanyze (2017), este último avaliando somente os um milhão maiores websites da internet.

Na Figura 5 temos alguns dados que ajudam a responder a última questão sobre a parcela do mercado atendida pelo *Google Analytics*.

Figura 5 - Participações no mercado de web analytics



Fonte: W3Techs (2017).

Figura 6 - Participações no mercado de web analytics nos um milhão maiores websites.

Posição	Tecnologia	Domínios	Parcela do Mercado
1	Google Analytics	647,975	42.01%
2	Google Universal Analytics	477,104	30.93%
3	Yandex Metrica	65,654	4.26%
4	LiveInternet	42,539	2.76%
5	comScore	40,079	2.6%
6	Quantcast	37,332	2.42%
7	Hotjar	24,189	1.57%
8	CrazyEgg	18,380	1.19%
9	Histats	16,984	1.1%
10	Piwik	16,307	1.06%

Fonte: Datanyze (2017).

Importante ressaltar que na Figura 6 deve-se somar a participação do *Google Analytics* e *Google Universal Analytics*, pois são versão diferentes para a mesma ferramenta de *web analytics* do Google Inc.

Com estes dados, concluímos que, ao realizar o estudo com base nesta ferramenta, estaremos abarcando em torno de pelo menos 83% dos websites que tem algum tipo de ferramenta de *web analytics*, tornando amplo nosso universo de oportunidade.

3.11 MÉTRICAS E KPIS

Assim como na escala dado, informação e conhecimento, métricas e *KPIs* também têm uma relação e são utilizadas como base para avaliações tanto quantitativas quanto qualitativas do negócio. Farris (2013) define métricas como o sistema de mensuração que quantifica uma tendência uma dinâmica ou característica. Sendo estas utilizadas na explicação de fenômenos, identificar causas, criando a possibilidade de comparação de tais observações em diferentes espaços de tempo. Portanto são somente as medidas geradas pela avaliação de um acontecimento, não tendo necessariamente relação com os objetivos da organização, sendo este seu principal ponto de diferenciação para com os indicadores-chave de performance, como afirmado por Rozner (2013) em seu relatório para a Agência dos Estados Unidos Para O Desenvolvimento Internacional (USAID), a definição mais simples que se pode ter de um *KPI* é: medidas que um setor ou organização utiliza para definir seu sucesso e avaliar o progresso na conquista de seus objetivos estratégicos.

Indicadores-chave de Performance são, como descritos por Popa (2015), elementos importantes para o atingimento de objetivos organizacionais, pois:

- Permitem a avaliação e determinação do progresso ao objetivo;
- Guiam a estratégia organizacional;
- São considerados expressões quantitativas e qualitativas da execução da estratégia.

3.11.1 Diferença entre Métrica e KPI

Iniciando-se pela definição inicial de Rozner, encontramos autores que elaboraram uma descrição mais profundo sobre o que diferencia um KPI, classificando-o como um subgrupo das métricas, como discorre Popa (2015), um *KPI* é uma métrica, mas uma métrica não necessariamente é um *KPI*. A relação contrária somente se dá como válida se uma métrica pode ser usada como fonte de análise que resulte em planos de ação de forma positiva. Desta forma reafirma-se a forte relação dos indicadores com os objetivos da organização realizando a análise.

Em posse destas afirmações é possível compreender uma característica marcante de um *KPI*, base para a diferenciação em relação às métricas, de acordo com Da Matta (2017, grifo da autora), *"KPIs* não são universais. O que serve para uma empresa pode não servir para outra. Para serem relevantes eles precisam refletir os objetivos de negócios.".

3.12 CARACTERÍSTICAS DE UM KPI

Mesmo que *KPIs* compartilhem fundamentalmente as mesmas características das métricas, este conjunto é estendido por mais alguns itens, como afirmado por Mortensen (2008), *KPIs* têm sete características particulares em relação às métricas, são elas:

- Remetem aos objetivos organizacionais;
- São determinados pela direção;
- Provêm contexto;
- Criam significado em todos os níveis da organização;
- São baseados em dados legítimos;
- Fáceis de compreender;
- Direcionam à ação.

Em sua publicação, Da Matta (2017, p. 8) descreve critérios para escolha de um bom *KPI* para o negócio digital. Segundo a autora, *KPIs* selecionados devem ter três atributos: simplicidade, relevância e rapidez. Descritos abaixo:

3.12.1 Simplicidade

Trata a complexidade da informação, no que diz respeito a facilidade de assimilação e no entendimento de seu significado aos envolvidos na análise, não necessitando demasiadas explicações.

3.12.2 Relevância

O *KPI* deve ser relevante, ou seja, devem ter relação direta com as regras e objetivos do negócio, "*KPIs* não são universais" (DA MATTA, 2017). Cada objeto analisado tem suas particularidades perante o mercado e com isto deve ser avaliado com o correto conjunto de informações.

3.12.3 Rapidez

Tange sobre a velocidade de obtenção da informação, se houver grande despendimento de tempo para a coleta da informação, isto pode demonstrar um possível impacto negativo quanto capacidade de resposta em momentos onde o tempo de resposta é crucial.

Outros autores, como Knezovic (2014) e Holman (2009), *KPIs* devem seguir o modelo SMART, ou seja, deve conter 5 características, descritas no Quadro 2.

Quadro 2 - Características de um KPI

Específico (*Specific*)	O que é avaliado deve estar claro a todas as pessoas envolvidas na avaliação, havendo entendimento comum.
Mensurável (*Measurable*)	Deve haver uma unidade de medida para o KPI, seja em moeda corrente, porcentagem ou numeral.
Alcançável (*Achievable*)	A meta definida para o KPI deve estar de acordo com a realidade do negócio analisado, sendo plausível sua alcançabilidade.
Relevante (*Relevant*)	O *KPI* deve prover perspectivas sobre o desempenho da organização.
Temporal (*Timely*)	Deve haver uma relação direta do indicador com o período de tempo analisado.

Fonte: Holman (2009).

Por fim, temos a definição sucinta de Zotos (2011) em sua publicação, onde *KPIs* referem-se a um conjunto de medições que refletem a performance ou sucesso de uma organização quanto ao progresso de seus objetivos. Segundo o autor, *KPIs* devem ter as seguintes características:

- Sejam indicadores de sucesso;
- Apresentados por meio de taxas;
- Requerem comparação;
- Dependem da indústria e do tipo de website.

Esta última característica indicada pelo autor a qual relaciona a relevância do indicador com o contexto onde o negócio está inserido é de suma importância para o atual trabalho, pois é esta relação um dos alicerces da organização deste trabalho.

3.13 CLASSIFICAÇÃO DE KPIS

Por terem relação direta com os objetivos organizacionais, *KPIs* tendem a ser particulares ao contexto do negócio e assim dificilmente tratados como globais, ou seja, *KPIs* compartilhados entre diversos tipos de negócios diferentes, pois, como referido por Kaushik (2007) *KPIs* definidos globalmente frequentemente não conseguem acomodar as diferenças estratégicas e processos de negócios, não se tornando assim tão úteis como acredita-se serem.

Por este motivo, é importante que na etapa de seleção e criação de indicadores, deve-se iniciar pelo principal objetivo a ser alcançado, para então a partir deste elaborar os objetivos secundários e finalmente, criar os *KPIs* referentes (POPA, 2015).

No contexto desta pesquisa, o valor semântico relativo às informações representadas pelos *KPIs* é uma peça fundamental para a classificação destes nos os possíveis conjuntos de análise. Este valor semântico pode variar de acordo com o tipo de negócio ou *website* analisado, existindo um grupo de *KPI* que poderiam ser considerados principais, que são de interesse para qualquer tipo de *website*, e outros que pode ou não ter valor, dependendo do mercado. Não há uma definição clara destes grupos, autores divergem quanto a suas classificações, listada abaixo o modelo de classificação de Zotos (2011) a título de ilustração das formas de classificação possíveis.

Zotos (2011), lista um esboço de classes para os *KPIs* de acordo com sua linha de raciocínio:

> Criar **metas mensuráveis** e específicas é um passo precursor vital para a definição de indicadores-chave de performance. Dependendo do seu tipo, um *website* pode ter objetivos completamente diferentes de outros. Objetivos comuns para um *E-commerce* são o aumento do número de compras, número de itens comprados e valor médio do *ticket*, enquanto para websites de conteúdo, estes objetivos são: aumento do consumo de conteúdo, número de inscritos, visualização de vídeos, número de jogadores on-line etc. (ZOTOS, 2011, tradução nossa, grifos do autor).

Em sua publicação, o autor agrega os *KPIs* em cinco grupos semânticos, porém não descrevendo suas relações com tipos de negócios digitais. Estes grupos são:

KPIs gerais sobre o website

- Taxa de conversão;
- Taxa de conversão de metas;
- Grupos de usuários;
- *Bounce Rate*;
- Tempo no *website*;
- Tipos de fonte de tráfego.

KPIs de visibilidade

- Tráfego de palavras-chave sem a marca;
- Tráfego gerado por termos específicos;
- *Bounce rate* por palavra-chave;
- *Rank* de palavra-chave;
- Visitantes novos e recorrentes.

KPIs de interação

- Interações de mídia social;
- Consumo de mídia;
- Contatos/Inscrições.

KPIs transacionais

- Custo por transação;
- Ticket media de transação;
- Média de itens no cesto de compras;
- Taxa de conversão por canal de aquisição.

KPIs geográficos

- Transações por país/região geográfica;
- *Bounce rate* por país/região geográfica;
- Distribuição de tráfego por país/região geográfica.

4 TRABALHOS RELACIONADOS

Neste capítulo são descritos trabalhos acadêmicos relacionados ao uso de classificação nos quais árvores de decisão são amplamente utilizadas em classificações de *data sets* nas mais diversas áreas de conhecimento a fim de criar artefatos de conhecimento sobre estes dados, mas não fazem uso todos os conceitos ou ferramentas utilizadas neste trabalho. Portanto listados abaixo estão trabalhos comerciais ou acadêmicos que se assemelham ao propósito deste trabalho.

4.1 UTILIZANDO TECNOLOGIAS DE WEB SEMÂNTICA E TÉCNICAS DE MINERAÇÃO DE DADOS PARA ANALISAR OS ESTUDANTES QUE APRENDEM E PREVER O DESEMPENHO FINAL

Sistemas de *e-learning* tem se tornado cada vez mais populares em ambientes de ensino como meio de distribuição de educação. Técnicas de *data mining* têm sido recentemente utilizadas por pesquisadores e professores para analisar o aprendizado de seus alunos com objetivo de entender e ter uma visão mais detalhada do processo, possibilitando melhorar a qualidade do ensino. Neste trabalho acadêmico é apresentada uma metodologia para analisar o aprendizado e extrair regras semânticas que podem ser usadas para predizer a performance do estudante ao final do curso. Isto é feito utilizando árvores de decisão para discriminar uma categoria especial de estudantes que correm risco de reprovação, para que estes recebam auxilio extra para melhoria do desempenho (GRIVOKOSTOPOULOU; PERIKOS; HATZILYGEROUDIS, 2014).

4.2 WEB SPAM DETECTION USING IMPROVED DECISION TREE CLASSIFICATION METHOD

O *web spam*, que significa a manipulação dos algoritmos de ranqueamento de mecanismos de busca por parte de website para obtenção de posições melhores do que os merecidos, atualmente se tornou uma séria preocupação para os buscadores da internet, os quais buscam melhores métodos para identificar a ocorrência de tal manipulação. São utilizadas técnicas, entre elas a classificação de características de websites utilizando o algoritmo C5.0, uma nova implementação do conhecido C4.5, buscando identificar elementos que determinem a manipulação, resultando no rebaixamento do ranking do website (TUNDALWAR; KULKARNI, 2014).

4.3 A NOVEL APPROACH FOR EFFECTIVE WEB PAGE CLASSIFICATION

Devido ao volume de dados cada vez maior disponível na internet, houve impacto no processo de classificação destes dados e dos websites que os contém, tornando os classificadores menos eficientes para o trabalho. Como resposta a isto, abriram-se duas linhas de trabalho, melhoria dos classificadores ou melhorar a parametrização dos dados que serão classificados. Sendo este último o foco do trabalho dos autores em seu artigo, onde propõem um método como alternativa para as opções tradicionais de classificadores (MANGAI; KUMAR; BALAMURUGAN, 2013).

4.4 SEMANTICS-BASED WEB SERVICE CLASSIFICATION USING MORPHOLOGICAL ANALYSIS AND ENSEMBLE LEARNING TECHNIQUES

Com a evolução da internet para um novo paradigma, *Web* de serviços, onde dados e serviços podem ser reutilizados entre aplicações em suas mais diversas formas e escala, este conjunto de serviços acabam por gerar um problema quanto a sua facilidade de ser listada e descoberta por desenvolvedores de aplicações de serviço. Como proposta para melhorar o cenário apresentado, é realizada um ensaio que estende o conceito de análise de similaridade de serviços por meio de vetorização e categorização das características semânticas dos serviços, possibilitando consequentemente sua classificação perante o conjunto e assim facilitando sua identificação e uso por desenvolvedores (KAMATH; ANANTHANARAYANA, 2016).

4.5 ANÁLISE COMPARATIVA

Cada um dos trabalhos relacionados apresenta formas eficazes de uso dos algoritmos de classificação em diversas áreas como uma ferramenta geradora de conhecimento. Coincidem com o trabalho apresentado etapas como ETL em alguns, ferramentas ou algoritmos de classificação em outros, porém todos estão centrados na criação de conhecimento por meio da classificação de dados. Este trabalho também demonstram a eficácia do uso do algoritmo classificador C4.5, sendo usado, ora efetivamente no trabalho, ora como benchmark para comparação de desempenho, corroborando assim o uso do algoritmo C4.5 neste trabalho.

No Quadro 3 estão categorizados os trabalhos de acordo com características relevantes a este trabalho: uso de *web analytics*; algoritmo de classificação utilizado; existência de uma etapa de extração, transformação e carregamento; ferramentas e/ou técnicas utilizadas, uso de indicadores-chave de desempenho.

Quadro 3 - Comparativo características de trabalhos relacionados

Índice	Web Analytics	Algoritmo de Classificação	ETL	Ferramentas Utilizadas	KPI
3.1	NÃO	C4.5 (J48), CART (SimpleCart)	NÃO	AITS, Weka, Protégé	SIM
3.2	NÃO	C4.5 (C5.0)	NÃO	KL Divergence, Open Directory Project	SIM
3.3	NÃO	C4.5, Bayes, ID3, oneR, MLP, kstar, SVM	SIM	Medida AVS, PWPC, WebKB, WEKA	NÃO
3.4	SIM	Multionomial Näive Bayes, SVM	SIM	NLP, XML, PCA	SIM
Trabalho	**SIM**	**C4.5 (J48)**	**SIM**	**GA, Pentaho, Weka**	**SIM**

Fonte: elaborado pelo autor.

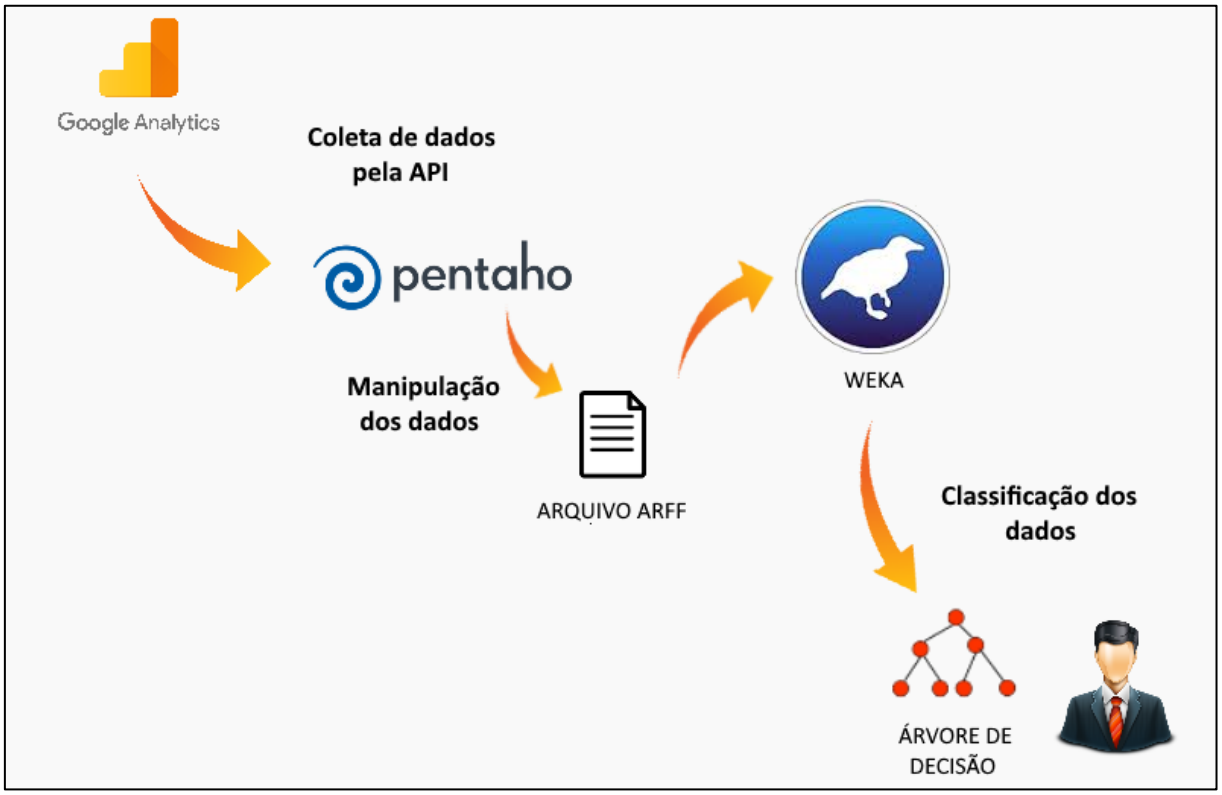
5 SOLUÇÃO PROPOSTA

Neste capítulo há a descrição da solução proposta para criação do conhecimento sobre o website. Sendo descrita a visão geral do sistema, a análise de requisitos funcionais e não-funcionais, modelagem da solução, regras de negócios e limitações impostas pelas ferramentas escolhidas.

5.1 VISÃO GERAL DO SISTEMA

A solução proposta coleta as métricas selecionadas através da API de desenvolvedores do *Google Analytics*, isto é feito utilizando criando uma conexão de serviço entre do Pentaho PDI, onde, após coletados, são pré-processados para que respeitem as características requeridas para a sua futura classificação. Os dados são então salvos no sistema de arquivo em formato ARFF – *Attribute-Relation File Format* – no qual é lido pelo programa de *data mining* Weka para sua classificação utilizando-se o algoritmo J48 para elaboração da árvore de decisão, que é o produto final do processo. Todas as macro-etapas do processo estão descritas na Figura 7.

Figura 7 - Visão geral do processo sugerido



Fonte: elaborada pelo autor.

Para a elaboração do processo estão sendo utilizadas a versão *Community* do Pentaho 7.0, API *Google Analytics* v3.0 em sua versão gratuita, Weka 3.8.1, JAVA 1.8.0 u131 64bits e como sistema operacional *Windows* 10 versão 1703 64bits.

5.2 REGRAS DE NEGÓCIOS

Com objetivo de definir quais são as regras de negócio e requisitos de uso para execução da solução proposta, serão listadas a seguir as ferramentas utilizadas, definindo seu propósito no processo.

Coleta de Dados pela API *Google Analytics*: fase inicial do processo, onde o *software* Pentaho conecta diretamente à API do *Google Analytics*, são definidas as configurações da propriedade em questão e escolhidas as métricas desejadas para a classificação, o período de avaliação e a granularidade dos dados para coleta das métricas solicitadas. Esta solicitação deve ser feita utilizando-se de uma conta de desenvolvedor devidamente habilitada e com identidade verificada no momento da conexão, isto sendo descrito nos requisitos funcionais desta etapa.

Manipulação dos dados Pentaho: logo após a coleta de dados, este são filtrados para remoção de dados complementares enviados pela API juntamente aos dados solicitados, pois não tem valor para a classificação posterior. Nesta etapa também são alteradas as tipificações de dados, enquadrando-os nos requisitos da etapa de classificação. Finalizando com a criação do arquivo ARFF, salvo no sistema de arquivos do sistema operacional.

Classificação dos dados Weka: após o carregamento do arquivo ARFF gerado pelo Pentaho na etapa anterior, no Weka é feita a preparação dos dados para sua classificação e configuração das opções do algoritmo J48, equivalente Weka do algoritmo de classificação C4.5, também são feitas as seleções de métricas a fim de aumentar a qualidade do resultado da classificação, resultando por fim, a árvore de decisão referente às métricas e a conversão escolhidas.

5.3 DETALHAMENTO DAS ETAPAS DO PROCESSO

Como pré-requisito à coleta das informações, é necessário que o *website*/aplicação avaliado tenha em seu código-fonte o *snippet* do *Google Analytics*, que permite a coleta das métricas de navegação durante a visita de usuários.

Quadro 4 - Snippet de monitoramento do *Google Analytics*

```
<script>
(function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(
){
(i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new
Date();a=s.createElement(o),
m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore
(a,m)
})(window,document,'script','https://www.google-
analytics.com/analytics.js','ga');

ga('create', 'UA-XXXXX-Y', 'auto');
ga('send', 'pageview');
</script>
```

Fonte: elaborado pelo autor.

Juntamente a este código deve-se ter evento específicos no website/aplicação, disparados durante a visitação ao website ou em situações específicas na aplicação. Para que seja possível realizar efetivamente a solução proposta neste trabalho deve haver ao menos um evento sendo monitorado. Estes eventos são monitorados por meio da chamada de uma função em *Javascript* ao *Google Analytics*, descrita no Quadro 5, quando uma situação relevante ocorre no website/aplicação.

Quadro 5 - Snippet de registro de evento do *Google Analytics*

```
ga('send', 'event', [eventCategory], [eventAction], [eventLabel],
[eventValue], [fieldsObject]);
```

Fonte: elaborado pelo autor.

Os parâmetros envoltos em colchetes são opcionais ao envio do registro de evento, porém cada um possibilita envio de informações adicionais quanto a ocorrência registrada, conforme imagem fornecida no guia da ferramenta.

Figura 8 - Parâmetros de registro de evento

Nome do campo	Tipo de valor	Obrigatório	Descrição
eventCategory	texto	sim	Normalmente, o objeto com o qual a interação ocorreu (por exemplo, 'Video')
eventAction	texto	sim	O tipo de interação (por exemplo, 'play')
eventLabel	texto	não	Útil para classificar eventos (por exemplo, 'Fall Campaign')
eventValue	número inteiro	não	Um valor numérico associado ao evento (por exemplo, 42)

Fonte: Google (2017).

Por se tratar de um processo para obtenção do conhecimento, suas etapas e seus passos serão descritos de acordo com as fontes e guias fornecidos pelos fabricantes dos *softwares* para realização das operações.

5.3.1 Coleta de dados pela API Google Analytics

Como principal pré-requisito para possibilitar a coleta de dados por meio da API do *Google Analytics*, deve-se primeiro ter uma conta Google que tenha acesso de administrador ao website analisado. Como primeiro passo temos a configuração da interface de desenvolvedor do GA, descritos a seguir:

- Acessar o website <http://console.developers.google.com/> e entrar com seus dados de acesso;
- Na página Projetos, criar um novo projeto, nomeá-lo e confirmar selecionando opção “criar”;
- No painel principal, selecionar a opção “Ativar API”;
- Procurar e selecionar a opção “Analytics API”, clicando em “Ativar API”;
- No menu lateral, selecione “Credenciais”, então clicar em “adicionar credencial”, escolhendo o tipo “Conta de serviço”;
- Quando perguntando sobre o tipo de chave, escolher o tipo P12 e então em “criar”.
- Salvar a chave P12 pois esta será utilizada no Pentaho para autenticar sua identidade futuramente no Pentaho;
- Copiar também o e-mail gerado para a conta de serviço, está será utilizada nos próximos passos;

- Acessar <http://google.com/analytics> utilizando sua conta administradora de um website analisado e selecione a opção “admin” no menu lateral;
- Adicionar permissão de leitura para o e-mail de serviço criado nos passos anterior ao website em questão, tendo este ao menos acesso de leitura aos dados.

Realizados estes passos o *Google Analytics* permitirá acesso aos dados do website por parte do Pentaho, cumprindo a primeira etapa do processo.

5.3.2 Manipulação dos dados Pentaho

O Pentaho deve ser configurado para que colete os dados da API e permita a transformação e consequente exportação de dados, o detalhamento abaixo abrange apenas as etapas minimamente necessárias para o funcionamento correto do processo, deixando possibilidade de transformações nos dados conforme as eventuais necessidades de análise por parte dos avaliadores dos dados.

- Abrir um novo projeto no Pentaho;
- No menu lateral, aba “Design”, pesquise pelo passo “Google Analytics”, do tipo Input. Insira-o com duplo-clique;
- Abrir as opções do passo com duplo clique no item inserido no painel do projeto;
- Insirir o e-mail da conta de serviço no campo “OAuth Service Email”;
- Insirir o local da chave P12 salva anteriormente;
- Selecionar qual website será avaliado na entrada “Get tableId from profile”, primeiro clicando em “Get profiles” e depois selecionando na lista atualizada;
- Escolher o período cronológico a ser pesquisado, inserindo as datas de início e fim, no formato ano-mês-dia;
- Selecionar as dimensões e métricas a serem coletadas no período escolhido;
- Antes de salvar as mudanças, solicitar os campos de saída do passo, clicando em “Obtem Campo” e em seguida “OK” para salvar o passo.

Para escolher as dimensões e métricas, deve ser utilizada sua nomenclatura técnica, descrita no link <https://developers.google.com/analytics/devguides/reporting/core/dimsmets> (em inglês). Para ser coletado o dado em sua menor granularidade possível, é sugerido a utilização da dimensão “ga:nthMinute” que lista as entradas minuto a minuto, juntamente a esta

dimensão deve ser adiciona às métricas “ga:goalCompletionsAll” que retorna a quantidade de conversões registrados nas entradas do banco. Outras métricas e dimensões podem ser escolhidas com objetivo de coletar informações sobre a visita, o visitante, dispositivo usado, entre outros.

Para exportar o arquivo ARFF, necessário para a próxima etapa do processo, será necessário adicionar um plugin ao Pentaho. Isto pode ser feito utilizando a própria interface do *software*. Para realizar os seguintes passos.

- Selecionar a opção “Tools” no menu superior, em seguida “Marketplace”;
- Insirir no filtro de busca o valor “Arff”, selecione o item filtrado e instale clicando no botão “Install”, confirme em “OK”;
- Salvar o projeto e reinicie o Pentaho para a instalação fazer efeito;
- No menu lateral, aba “Design”, pesquise pelo passo “Arff Output”, do tipo data mining. Insira-o com duplo-clique;
- Abrir as opções do passo com duplo clique no novo item inserido no painel do projeto;
- Selecionar o caminho e nome do arquivo a ser salvo no sistema de arquivos na opção “File name”;
- Selecionar a aba “Fields” para configurar a transcrição de tipos do Kettle para o tipo ARFF. Clique em “Obtem campos”, selecione a opção “Nominal” para cada entrada do tipo String, “Numeric” para cada entrada do tipo “Integer” ou “Number”;
- Confirmar as configurações clicando em “OK”;
- Conectar dos dois passos com um *hop,* a partir do passo Google Analytics e chegando ao passo Arff Output;
- Rodar a transformação do Pentaho selecionando “Action” no menu superior e em seguida “Run”, confirmando o modal aberto clicando em “Run”.

Há a possibilidade do console da aplicação apresentar um erro ao final da execução dos passos, porém após verificação da integridade dos dados gerados e salvos no sistema de arquivos, constata-se que o erro não impactou o *output* dos passos, ocorrendo somente ao fim da execução.

5.3.3 Classificação dos dados Weka

Última etapa do processo, foca exclusivamente na aplicação do método de classificação dos dados, juntamente com as operações de preparo prévio dos dados. Nesta proposta será utilizado o algoritmo de aprendizado J48, que é a implementação do algoritmo C4.5, muito referido em artigos e publicações acadêmicas sobre classificação de dados. Esta escolha foi feita por se tratar de um algoritmo conhecido no meio acadêmico e com compatibilidade de processamento da base de dados coleta do *Google Analytics*, porém não há impeditivo para uso de outros algoritmos de aprendizado para criação de árvores de decisão.

Os passos para a realização da classificação utilizando o Weka, listados a seguir:

- Selecionar a opção *Explorer*;
- Na aba "*Preprocess*" abrir o arquivo ARFF selecionando o botão "Open File...";
- Selecionar as colunas a serem utilizadas na classificação, é recomendado que a coluna que representa a granularidade de agregação das entradas do Google Analytics seja removida pois podem influenciar negativamente na qualidade da classificação. Nomes destas colunas podem ser "ga:nthMinute", "ga:nthHour" ou "ga:nthDay";
- Terminado o pré-processamento, selecionar a aba "Classify";
- Selecionar o algoritmo de aprendizado clicando no botão "Choose" e escolha "classifiers > tree > J48";
- Para opções de teste, serão utilizadas as configurações padrão, "Cross-validation" com 10 *Folds*;
- Selecionar o item "(Num) ga:totalEvents" como classe a ser avaliada;
- Clicar em "Start" para iniciar a classificação.

A ferramenta Weka permite que sejam realizadas operações de pré-processamento nos dados antes da execução da classificação, com objetivo de melhorar a qualidade dos resultados obtidos, entretanto estas operações não são cobertas por esta proposta.

Seguindo estes passos, sendo as métricas e dimensões selecionadas as descritas neste detalhamento, a classificação retornará uma árvore de decisão podada que representa do modelo de predição, determinando que aspectos são mais importantes para a determinar se uma visita resultará em um evento ou uma conversão no website/aplicação avaliado.

5.4 EXPERIMENTOS REALIZADOS

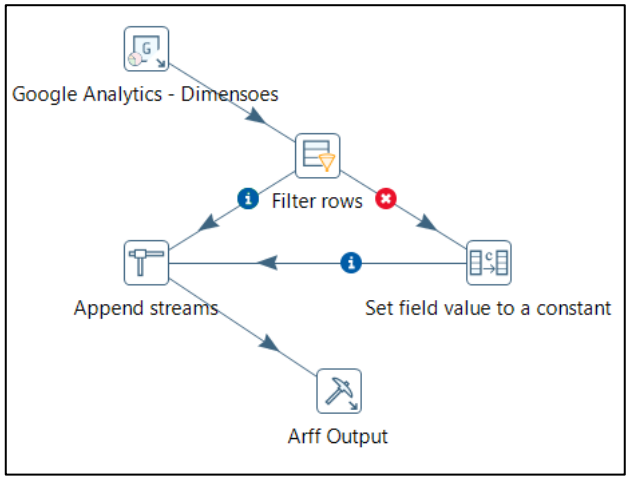
Para verificar a eficácia do processo proposto, foram realizados dois experimentos utilizando a configuração e etapas descritas no item 4.3 deste trabalho. Foram selecionados dois websites institucionais de empresas, uma escola e uma fundação que utilizam a internet para realização de seus negócios, denominados *Website* 1 e *Website* 2, mantendo sua identidade sigilosa, a pedido de seus gestores.

5.4.1 Parâmetros dos experimentos

Para a avaliação de ambos *websites* foram selecionadas treze métricas, sendo uma delas o atributo classe "ga:goalCompletionsAll". As restantes são: mídia de acesso (ga:medium), hora de acesso (ga:hour), tipo do dispositivo utilizado para acesso (ga:deviceCategory), categoria de usuário (ga:userType), dia da semana do acesso (ga:dayOfWeek), dias desde a última visita (ga:daysSinceLastSession), número de visitas (ga:sessions), rejeições (ga:bounces), visualizações de página (ga:pageViews), tempo na página (ga:timeOnPage), duração da visita (ga:sessionDuration) e interações (ga:hits).

Durante a etapa de manipulação dos dados no Pentaho, o atributo de número de eventos/conversões do período foi normalizado em 0 ou 1, ocorrência ou não-ocorrência do evento, para isto foram utilizadas etapas de transformação de dados que separaram os registros com a quantidade de evento maiores que 0 para que estes tenham seu valor definido como 1 e por fim unidos novamente com o restante dos registros. Tal transformação foi organizada conforme a Figura 9.

Figura 9 - Transformação de dados no Pentaho



Fonte: elaborada pelo autor.

Para a classificação, foi removido o atributo identificador do registro (ga:nthMinute), como sugerido pela bibliografia relacionada, utilizada a implementação J48 do algoritmo C4.5 com suas configurações padrão, fator coincidência de 0.25, número mínimo de objetos por folha de 2 e 10-fold cross-validation.

5.4.2 Experimento Website 1 - Escola

Para avaliar o *website* da escola, foi selecionado o período de 30 dias anteriores à data de realização do teste, de 16 de abril de 2017 à 17 de maio de 2017 que, ao serem agregados por minuto, resultou em 9.225 registros. Dentre estes registros, 4.373 registros foram identificados como ocorrência de evento/conversão, 47,4% do total.

5.4.3 Experimento Website 2 - Fundação

Para avaliar o website da escola, foi selecionado o período de em que se iniciou o monitoramento de eventos em seu novo *website*, 9 de maio de 2017 até a data de realização do teste, 18 de maio de 2017, resultando em 10 dias de dados que, ao serem agregados por minuto, resultou em 5503 registros. Dentre estes registros, 68 registros foram identificados como ocorrência de evento/conversão, 1,02% do total.

6 RESULTADOS

As árvores de decisão resultantes, demonstraram uma variação em seus indicadores de precisão, onde uma oscilou próximo a 99%, já outra próxima de 80%, fato que levanta perguntas sobre uma possível relação de dependência entre o resultado da classificação com a arquitetura de informação do website, ou seja, a forma como seu conteúdo é organizado e distribuído em suas páginas, fato a ser verificado em trabalhos futuros. Não se sabe exatamente quais as características desta dependência e qual seu impacto sobre as métricas, ficando como exemplo mais perceptível a métrica de duração da visita, em que em grande parte dos registros teve sua duração de visita (ga:sessionDuration) nula. Após investigação, foi determinado que tal situação ocorre devido à forma que o Google Analytics calcula e cria a métrica, calculando a diferença entre o momento de ocorrência de duas interações com o servidor do GA. Em websites e aplicações que contém apenas uma página, por exemplo, esta métrica tende a permanecer zerada, o que impacta negativamente na qualidade da classificação que a utiliza.

6.1 EXPERIMENTO WEBSITE 1 – ESCOLA

O número de registros classificados de forma correta no experimento foi de 9.188, representando uma Acurácia de 99,59%, já a precisão alcançada foi de 99,72%, com 4.361 positivos-verdadeiros e 12 falsos-negativos. Outras características estão descritas no Quadro 6, que apresenta o *output* bruto do Weka.

Quadro 6 - Resultado bruto da classificação “Website 01”

(continua)

```
=== Run information ===

Scheme:   weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:  NewRelation-weka.filters.unsupervised.attribute.Remove-R1
Instances:  9225
Attributes: 13
       ga:medium
       ga:deviceCategory
       ga:userType
       ga:hour
       ga:dayOfWeek
       ga:daysSinceLastSession
       ga:sessions
       ga:totalEvents
       ga:bounces
       ga:pageViews
       ga:timeOnPage
       ga:sessionDuration
       ga:hits
```

Quadro 6 - Resultado bruto da classificação “Website 01”

(continua)

```
Test mode:  10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

ga:hits <= 3
| ga:pageViews <= 0: 1 (62.0/2.0)
| ga:pageViews > 0
| | ga:hits <= 1: 0 (3335.0)
| | ga:hits > 1
| | | ga:pageViews <= 1
| | | | ga:bounces <= 0
| | | | | ga:timeOnPage <= 6: 1 (24.0/5.0)
| | | | | ga:timeOnPage > 6: 0 (24.0/10.0)
| | | | ga:bounces > 0: 0 (4.0)
| | | ga:pageViews > 1
| | | | ga:hits <= 2: 0 (1198.0)
| | | | ga:hits > 2
| | | | | ga:pageViews <= 2: 1 (12.0/2.0)
| | | | | ga:pageViews > 2: 0 (200.0)
ga:hits > 3
| ga:pageViews <= 3
| | ga:hits <= 4
| | | ga:sessionDuration <= 53: 1 (621.0/1.0)
| | | ga:sessionDuration > 53
| | | | ga:pageViews <= 1: 1 (51.0)
| | | | ga:pageViews > 1: 0 (7.0)
| | ga:hits > 4: 1 (3129.0/1.0)
| ga:pageViews > 3
| | ga:hits <= 6: 0 (79.0/1.0)
| | ga:hits > 6
| | | ga:pageViews <= 6: 1 (418.0)
| | | ga:pageViews > 6
| | | | ga:hits <= 8: 0 (5.0)
| | | | ga:hits > 8: 1 (56.0)

Number of Leaves :    16

Size of the tree :    31


Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances    9188       99.5989 %
Incorrectly Classified Instances    37        0.4011 %
Kappa statistic             0.992
Mean absolute error           0.0047
Root mean squared error         0.0537
Relative absolute error         0.9403 %
Root relative squared error      10.7642 %
Total Number of Instances      9225
```

Quadro 6 - Resultado bruto da classificação “Website 01”

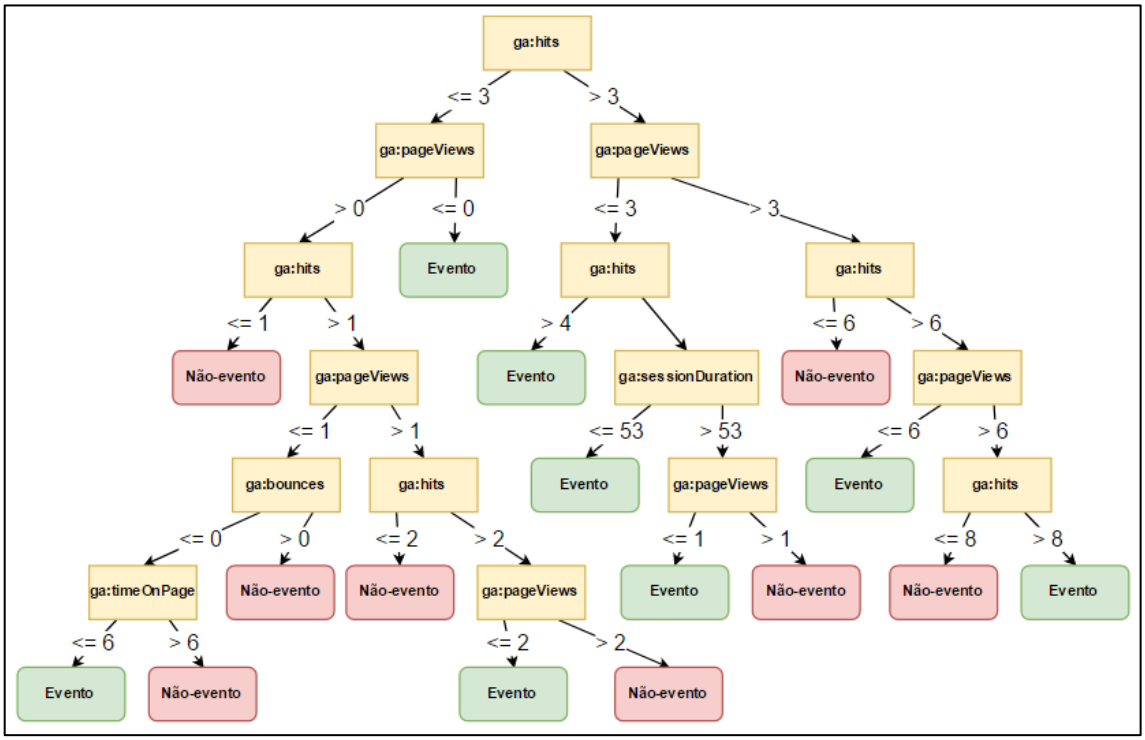
(conclusão)

```
=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0,995   0,003   0,998    0,995   0,996    0,992  0,999    0,998    0
0,997   0,005   0,994    0,997   0,996    0,992  0,999    0,998    1
Weighted Avg.
0,996   0,004   0,996    0,996   0,996    0,992  0,999    0,998

=== Confusion Matrix ===

   a    b <-- classified as
 4827   25 |   a = 0
  12 4361 |   b = 1
```

Fonte: elaborado pelo autor.

Sendo convertido o resultado para uma representação gráfica – Figura 10 – para melhorar sua legibilidade, como demonstrada na fundamentação teórica, sendo representado pela seguinte árvore de decisão.

Figura 10 - Árvore de decisão “Website 01”



Fonte: elaborada pelo autor.

Em sua forma podada, resultante do algoritmo J48, esta árvore tem tamanho – quantidade de nodos e folhas – 31 e 16 folhas, descrevendo a ocorrência ou não-ocorrência do evento analisado. Não apresenta uma legibilidade tão facilitada quanto esperado, porém é

possível compreender a dinâmica entre as métricas e sua influência no resultado final, na classe-folha da árvore.

6.2 EXPERIMENTO WEBSITE 2 – FUNDAÇÃO

O número de registros classificados de forma correta neste experimento foi de 5.482, representando uma acurácia de 99,61%, ponderada, já a precisão alcançada foi de 79,41%, com 54 positivos-verdadeiros e 14 falsos-negativos. Outras características estão descritas no Quadro 7, que apresenta o *output* bruto do Weka.

Quadro 7 - Resultado bruto da classificação "Website 02"

(continua)

```
=== Run information ===

Scheme:   weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:  NewRelation-weka.filters.unsupervised.attribute.Remove-R1,15-19
Instances:  5503
Attributes: 13
       ga:medium
       ga:deviceCategory
       ga:userType
       ga:hour
       ga:dayOfWeek
       ga:daysSinceLastSession
       ga:sessions
       ga:totalEvents
       ga:bounces
       ga:pageViews
       ga:timeOnPage
       ga:sessionDuration
       ga:hits
Test mode:  10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

ga:pageViews <= 0: 1 (41.0/2.0)
ga:pageViews > 0
| ga:hits <= 1: 0 (3712.0)
| ga:hits > 1
| | ga:pageViews <= 1
| | | ga:userType = New Visitor: 1 (21.0/5.0)
| | | ga:userType = Returning Visitor: 0 (29.0/1.0)
| | ga:pageViews > 1: 0 (1700.0/12.0)

Number of Leaves :      5

Size of the tree :      9

Time taken to build model: 0.02 seconds
```

Quadro 7 - Resultado bruto da classificação “Website 02”

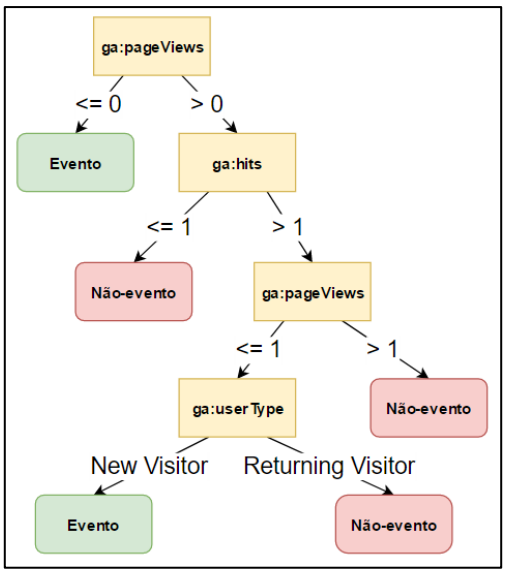
(conclusão)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        5482               99.6184 %
Incorrectly Classified Instances        21                0.3816 %
Kappa statistic                          0.8353
Mean absolute error                      0.0069
Root mean squared error                  0.0605
Relative absolute error                 28.0286 %
Root relative squared error             54.7324 %
Total Number of Instances             5503

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               0,999    0,206    0,997     0,999    0,998      0,837    0,943     0,999     0
               0,794    0,001    0,885     0,794    0,837      0,837    0,943     0,723     1
Weighted Avg.
               0,996    0,203    0,996     0,996    0,996      0,837    0,943     0,996

=== Confusion Matrix ===

   a    b   <-- classified as
 5428    7 |    a = 0
  14   54 |    b = 1
```

Fonte: elaborado pelo autor.

Sendo convertido o resultado para melhorar sua legibilidade, pode ser gerada uma representação gráfica, como demonstrada na fundamentação teórica, a classificação resultante é representada pela seguinte árvore de decisão.

Figura 11 - Árvore de decisão “Website 02”



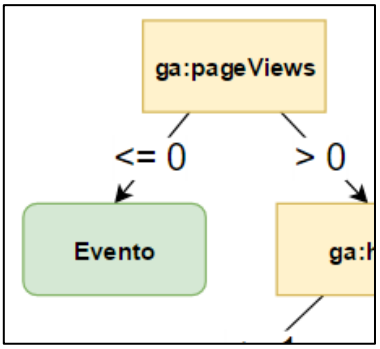
Fonte: elaborada pelo autor.

A árvore de menor tamanho, 9 e somente 5 folhas, sendo consideravelmente menor que a árvore do experimento número 01, possivelmente influenciado pela quantidade inferior de registros de eventos existentes na base de dados, mas com alta legibilidade.

6.3 DISCUSSÃO DOS EXPERIMENTOS

Percebe-se com estes dois experimentos, dois resultados finais distintos, sendo no experimento do Website 01, que a árvore apresenta um tamanho próximo do esperado pelo autor. No experimento Website 02, temos a aparição de uma variável binária, mostrando que estas podem contribuir ativamente para a ocorrência da classe evento. Porém, por mais que as árvores de decisão apresentam indicadores de precisão e acurácia aceitáveis, ao ser feita uma análise mais minuciosa quanto aos valores de atributos selecionados como nodos da árvore, como por exemplo na Figura 12, vemos situações que não condizem com uma situação real, onde com número de visualização de página (*pageviews*) menor ou igual a zero, resulta-se em evento.

Figura 12 - Detalhe da árvore de decisão referente ao "Website 02"



Fonte: elaborada pelo autor.

Avaliando se esta característica advém dos dados coletados ou se é uma anomalia resultante do processo de coleta e transformação dos dados, percebeu-se que a informação consta já na base de dados do Google Analytics. Este fato levantou algumas dúvidas:

- Seria o Google Analytics uma ferramenta eficaz para esta etapa do processo? Quais suas limitações?
- Há alguma característica dos websites avaliados que resultam nesta situação? Qual o impacto que a estrutura do website pode ter nas métricas que o Google Analytics coleta?

Para a primeira dúvida, será necessário aprofundar o estudo para elucidar a forma de obtenção destas métricas tidas como anômalas, para assim entender o que pode causa-las com mais precisão. Sendo o Google Analytics uma ferramenta utilizada intensivamente pelo mercado mundial para medição de desempenho de websites, pode-se descartar a possibilidade que seja algum problema estrutural na aplicação, o que nos leva à segunda dúvida, dando foco nos objetos monitorados pelo Google Analytics, terão estes websites alguma característica que causa tais anomalias nas métricas? No caso do Website 02, sabe-se que o evento avaliado ocorre na página principal do website, já no Website 01, este evento só é realizado em uma subpágina do website. Junto a isto temos o volume de eventos registrados, assim como a seleção de métricas a serem utilizadas na criação da árvore de decisão, estes podem ser fatores determinantes na qualidade do conhecimento gerado ao fim do processo. Para chegar a uma conclusão assertiva sobre isto será necessário um estudo específico para obtenção de resultados, ficando estas sugestões para trabalho posteriores nesta linha de estudo.

7 CONSIDERAÇÕES FINAIS

A solução proposta neste trabalho gerou, com tempo e complexidades relativamente baixos, em torno de 25 minutos, o artefato desejado em sua motivação, demonstrando assim sua eficácia. Nela estão envolvidos Pentaho, Weka e *Google Analytics*, *softwares* muito comuns na academia e no mercado de análise de dados e *Business Intelligence*, dando-se suas escolhas com base em sua aceitação, gratuidade, popularidade e oferta de documentação, estando os três em excelente posição em todos os quesitos.

Para a coleta de dados do *web analytics*, o *Google Analytics* se provou extremamente eficiente, embora tenha sido utilizada para a realização do trabalho sua versão 3, hoje já sendo ofertada a versão 4 da API. Esta escolha se deu por conta da compatibilidade do *software* Pentaho somente com a versão anterior, futuramente pode ser possível adaptar a metodologia do processo para utilizar a API em sua versão mais nova.

Quanto à manipulação dos dados, mesmo que a utilização do *software* Pentaho possa ser considerada baixa, a escolha de seu uso neste modelo mínimo do processo é baseada em seu potencial de transformação de dados, provendo assim ao gestor que utilizar o processo liberdade suficiente para fazer suas próprias transformações para geração de conhecimento de qualidade para seu contexto em particular. Para a exportação, poderiam ter sido utilizados outras formas de exportação de dados, mas o uso do plugin ARFF se provou a melhor opção por não ser necessária quaisquer alterações no arquivo após sua exportação, reduzindo a complexidade do processo ao não envolver outras ferramentas, como Excel.

Para a classificação, houve algumas dificuldades para adequar a base de registros ao algoritmo J48, somente após um processo de eliminação minucioso, destacou-se a necessidade do atributo-classe dos registros seja do tipo Nominal, o que acarretou em um ajuste no processo de manipulação, onde o atributo *ga:totalEvents* foi parametrizado para ter apenas dois valores possíveis, 0 ou 1, marcando sua ocorrência e não-ocorrência. Resolvido este caso, o restante do processo se provou simples e direto.

Nos ensaios realizados não foi avaliada a qualidade do artefato de conhecimento gerado ao fim do processo, focou-se totalmente na viabilidade de se conseguir o artefato, postergando o julgamento de qualidade para futuros trabalhos sobre o tema.

Conclui-se por fim com os resultados obtidos nos experimentos que é possível criar este artefato de conhecimento para dar suporte à decisão utilizando-se somente de ferramentas gratuitas e sem a necessidade de profundo conhecimento em BI ou *data mining*, possibilitando

o fornecimento de um tipo de conhecimento de fácil legibilidade para gestores de negócios que não são especialistas em tecnologia ou ciência de dados.

7.1 SUGESTÕES PARA TRABALHOS FUTUROS

O método proposto neste trabalho se prova eficaz para a criação do artefato de conhecimento pretendido para auxiliar gestores iniciantes na atuação de data mining e BI para seus negócios, porém sua versão apresentada podem receber melhorias para potencializar sua eficiência e qualidade do conhecimento gerado. Entre as melhorias possíveis, as listadas a seguir podem ser de grande valia:

- **Ajustes das variáveis de configuração do J48:** dependendo do website avaliado, suas métricas e suas características poderão responder melhor às configurações diferentes do J48, possibilitando assim mais qualidade no conhecimento gerado;
- **Utilização de outro algoritmo de classificação:** J48 é sugerido como algoritmo de aprendizado para a classificação por ser o algoritmo mais simples que atende às características dos dados providos pelo Google Analytics, mas outros poderão resultar em melhores conclusões;
- **Testar outras métricas e dimensões e avaliar AD resultante:** a ferramenta Google Analytics disponibiliza mais de 50 métricas e dimensões sobre o website/aplicação analisado, buscar formas de coletar todos estes dados respeitando os limites da API dará mais opções de avaliação por parte do gestor e também ajuda a entender o impacto na qualidade da árvore gerada;
- **Separação da classificação por tipo de evento:** como é possível registrar diferentes tipos de eventos e conversões no mesmo website, filtrar os dados para classifica-los em grupos diferentes dará um novo grau de detalhe ao conhecimento gerado;
- **Melhorias visuais:** tornar o resultado final visualmente agradável, trabalhando o resultado do processo utilizando Tableau, gerando gráficos e com isto melhorando sua legibilidade e aceitação entre gestores não-técnicos;
- **Aplicação Web:** disponibilizar uma versão baseada na web para facilitar seu uso, descartando a necessidade de o gestor aprender o funcionamento das ferramentas utilizadas no processo;

- **Testar a qualidade do conhecimento gerado:** entender como os gestores podem utilizar este conhecimento e adequá-lo ao contexto de uso, tornando-o mais útil e relevante aos gestores;
- **Como o Google Analytics coleta as métricas utilizadas:** saber exatamente como os dados são gerados pode trazer melhor conhecimento sobre características do website e como métricas de baixa qualidade podem ser evitadas;
- **Avaliar a relação entre estrutura do website e os dados coletados pelo Google Analytics:** para compreender como a arquitetura de informação do website pode influenciar as métricas coletadas pelo web analytic e consequentemente impacta na árvore de decisão gerada.

Diversos ajustes e operações de transformação de dados podem ser aplicadas durante a etapa de manipulação dos dados no Pentaho, estando isto a critério e criatividade do profissional que realizar o método proposto neste trabalho. A análise de informações de negócios deve sempre trazer perspectivas promissoras aos seus gestores e este é apenas o primeiro passo em direção a concretização deste objetivo.

REFERÊNCIAS

ANDRIOTTI, F. **A intuição no processo de tomada de decisão instantânea**. 2012. Tese (Doutorado em Administração) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

BATTITI, Roberto; BRUNATO, Mauro. **Reative Business Intelligence:** from data to models to insight. Reactive Research, 2011.

BERSON, Alex; SMITH, Stephen J. **Building data mining applications for CRM**. New York: McGraw-Hill, Inc., 2002.

BEYNON-DAVIES, Paul. **E-Business.** Hampshire: Palgrave Macmillan, 2004. 350 p.

BITTENCOURT, Anelise Caon. **Escuta permanente de informação informal e sua exploração coletiva para tomada de decisão:** uma observação participante na Johnson & Johnson un sul. 2013. 99f. Dissertação (Mestrado) – Curso de Administração, Pós-Graduação em Administração, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013. Disponível em: <http://www.ufrgs.br/gianti/files/orientacao/mestrado/defesa/pdf/AneliseCaondissert.pdf>. Acesso em: 9 set. 2015.

BOGORNY, Vânia. **Classificação:** conceitos básicos e árvores de decisão. Florianópolis: UFSC, 2015. 56 slides, color. Disponível em: <http://www.inf.ufsc.br/~vania/teaching/INE5644/classificacao_arvores.ppt>. Acesso em: 16 maio 2017.

CARLINI JUNIOR, Reginaldo José; VITAL, Tales Wanderley. A utilização do benchmarking na elaboração do planejamento estratégico: uma importante ferramenta para a maximização da competitividade organizacional. **Revista Brasileira de Gestão de Negócios: FECAP,** São Paulo, v. 1, n. 14, p. 60-66, abr. 2004.

CARVALHO, Marcelo Sávio Revoredo Menezes de. **A trajetória da internet no Brasil:** do surgimento das redes de computadores à instituição dos mecanismos de governança. 2006. 240f. Dissertação (Mestrado) – Curso de Engenharia de Sistemas e Computação, Coppe, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006. Disponível em: <https://www.researchgate.net/profile/Marcelo_Carvalho17/publication/268809917_A_TRAJETORIA_DA_INTERNET_NO_BRASIL_DO_SURGIMENTO_DAS_REDES_DE_COMPUTADORES_A_INSTITUICAO_DOS_MECANISMOS_DE_GOVERNANCA/links/54774a430cf2a961e4825bd4.pdf>. Acesso em: 21 mar. 2017.

CLIFTON, Brian. **Should you pay $150,000 for Google Analytics Premium?** 2015. Disponível em: <https://brianclifton.com/blog/2015/10/27/should-you-pay-150000-for-google-analytics-premium/>. Acesso em: 23 abr. 2017.

CLIFTON, Christopher. **Encyclopedia Britannica:** definition of data mining. Disponível em: <http://global.britannica.com/EBchecked/topic/1056150/data-mining>. Acesso em: 16 maio 2017.

DA MATTA, Norma Paiva. Universidade Estácio de Sá. **Métricas e monitoramento na web.** Disponível em: <http://pos.estacio.webaula.com.br/Cursos/POS571/docs/Aula_02.pdf>. Acesso em: 27 mar. 2017.

DATANYZE. **Analytics Market Share Report:** competitor analysis in Alexa top 1M. Disponível em: <https://www.datanyze.com/market-share/analytics/Alexa top 1M/>. Acesso em: 19 mar. 2017.

FARRIS, Paul W. et al. **Métricas de marketing**: o guia definitivo de avaliação de desempenho do marketing. 2. ed. Porto Alegre: Bookman, 2013. 426p.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 37-54, out. 1996

FERNANDES, Bruno G.; ROSA, Celso O. As métricas do marketing no século XXI. **Revista Panorama**, Goiás, v. 3, n. 1, p. 180, jan./dez. 2013. Disponível em: <http://estudos.ucg.br/index.php/panorama/article/download/3435/2006>. Acesso em: 22 jun. 2015.

FRIED, Jason; HANSSON, David Heinemeier. **Reinvente sua empresa:** mude sua maneira de trabalhar. Rio de Janeiro: Sextante, 2012.

GANGADHARAN, G. R.; SUNDARAVALLI, N. Swami. Business intelligence systems: design and implementation strategies. Information Technology Interfaces. INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY INTERFACES ITI, 23., 2004, **Anais…** jun. 2004.

GOOGLE INC. **Acompanhamento de eventos.** 2017. Disponível em: <https://developers.google.com/analytics/devguides/collection/analyticsjs/events>. Acesso em: 17 maio 2017.

GOOGLE INC. **Google Analytics:** features. Disponível em: <https://www.google.com/intl/en_ALL/analytics/features/>. Acesso em: 19 mar. 2017.

GOOGLE INC. **Sobre o comparativo de mercado:** compare o desempenho da sua propriedade com o dos seus colegas de setor. 2017. Disponível em: <https://support.google.com/analytics/answer/6086666?hl=pt-BR>. Acesso em: 21 mar. 2017.

GRIMES, Seth. **Unstructured data and the 80 percent rule.** Ago. 2010. Disponível em: <https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/>. Acesso em: 29 jun. 2017.

GRISE, M. L.; GALLUPE, R. B. Information overload in face-to-face electronic meetings: an integrative complexity approach. **Journal of Management Information Systems**, n. 16, p. 157-185, 1999.

GRIVOKOSTOPOULOU, Foteini; PERIKOS, Isidoros; HATZILYGEROUDIS, Ioannis. Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance. In: 2014 INTERNATIONAL CONFERENCE OF TEACHING, ASSESSMENT AND LEARNING (TALE), 3., 2014, Wellington. **Anais...** Wellington: Ieee, 2014. p. 488-494.

GROTH, R. **Data mining.** Englewood Cliffs: Prentice Hall, Inc., 1998.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning:** data mining, inference and prediction. 2. ed. Stanford: Springer, 2009. 745p.

HOLMAN, Victor. **Developing metrics and KPI (key performance indicators).** 2009. Disponível em: <https://pt.slideshare.net/victorholman/developing-metrics-that-drive-performance-success>. Acesso em: 27 mar. 2017.

HOUAISS, Antônio. **Dicionário Houaiss da Língua Portuguesa**. Rio de Janeiro: Ed. Objetiva, 2001.

HRONEC, S. M. **Sinais vitais**: usando medidas de desempenho da qualidade, tempos e custos para traçar a rota para o futuro de sua empresa. Arthur Andersen. São Paulo: Makron Books, 1994.

INTERNET LIVE STATS. **Total number of websites.** 2016. Disponível em: <http://www.internetlivestats.com/total-number-of-websites/>. Acesso em: 19 mar. 2017.

KAMATH, S. Sowmya; ANANTHANARAYANA, V. S. Semantics-based web service classification using morphological analysis and ensemble learning techniques. International **Journal of Data Science and Analytics**, [s.l.], v. 2, n. 1-2, p.61-74, 18 out. 2016. Springer Nature. http://dx.doi.org/10.1007/s41060-016-0026-x.

KAUSHIK, Avinash. **Web analytics:** an hour a day. Indianapolis: Wiley Publishing, 2007. 443p.

KNEZOVIC, Bojan. **Una mirada al futuro de métricas, KPIs & dashboards.** 2014. Disponível em: <https://pt.slideshare.net/BojanKnezovic/una-mirada-al-futuro-de-mtricas-kp-is-dashboards>. Acesso em: 27 mar. 2017.

KOBIELUS, James. What's Not BI? Oh, Don't Get Me Started.... Oops Too Late... Here Goes.... Abr. 2010. Disponível em: <http://blogs.forrester.com/james_kobielus/10-04-30-what%E2%80%99s_not_bi_oh_don%E2%80%99t_get_me_startedoops_too_latehere_goes>. Acesso em: 29 jun. 2017.

KUMARI, Navita. Business intelligence in a nutshell. **International Journal of Innovative Research in Computer and Communication Engineering**, Chennai, India, p. 969-975. jun. 2013. Disponível em: <https://www.ijircce.com/>. Acesso em: 7 abr. 2017.

LISBOA, Ruben. **Importância do Web Analytics para o seu negócio online.** 2012. Disponível em: <https://marketingdigitalpt.wordpress.com/2012/09/03/importancia-web-analytics-negocio-online/>. Acesso em: 9 out. 2015.

MAIA, Ana Paula de Assis et al. A decision-tree-based model for evaluating the thermal comfort of horses. **Scientia Agrícola**, Piracicaba, v. 70, n. 6, p. 377-383, dez. 2013. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-90162013000600001&lng=en&nrm=iso>. Acesso em 16 maio 2017.

MANGAI, J. Alamelu; KUMAR, V. Santhosh; BALAMURUGAN, S. Appavu. A novel approach for effective web page classification. **International Journal of Data Mining, Modelling and Management,** v. 5, n. 3, p. 233-245, mar. 2013.

MEHTA, Hrucha. **Learn about different types of web analytics.** 2015. Disponível em: <http://www.digitalvidya.com/blog/learn-about-different-types-of-web-analytics/>. Acesso em: 19 mar. 2017.

MITCHELL, Tom Michael. **Machine learning.** The McGraw-Hill Companies, Inc., 1997. 414p.

MORTENSEN, Dennis. **The difference between a KPI and a Metric.** New York: Visual Revenue, 2008. 25 slides, color. Disponível em: <https://www.slideshare.net/dennis.mortensen/the-difference-between-a-kpi-and-a-metric>. Acesso em: 9 abr. 2017.

NAGASUNDARAM, M.; DENNIS, A. R. When a group is not a group: the cognitive foundation of group idea generation. **Small Group Research**, n. 24, p. 463-489, 1993.

NONAKA, I.; TAKEUCHI, H. **Criação de conhecimento na empresa**: como as empresas japonesas geram a dinâmica da inovação. 12. ed. Rio de Janeiro: Elsevier, 1997.

O'BRIEN, James; MARAKAS, George. **Management Information Systems.** New York: Mcgraw-Hill/Irwin, 2011.

PINHEIRO, Marta M. K.; BRITO, Vladimir de P. Em busca do significado da desinformação. **Revista da Informação**, Minas Gerais, v. 15, n. 6, dez. 2014. Disponível em: <http://dgz.org.br/dez14/Art_05.htm>. Acesso em: 9 set. 2015.

POPA, Brînduşa Maria. Challenges when developing performance indicators. **Journal of Defense Resources Management (JoDRM),** Brasov, Romania, n. 1, ano 6, p. 111-114, out. 2015. Disponível em: <https://www.ceeol.com/search/article-detail?id=305960>. Acesso em: 8 abr. 2017.

POWER, D. J. **A brief history of decision support systems.** DSSResources.COM, v. 4.0, Mar. 2007. Disponível em: <http://DSSResources.COM/history/dsshistory.html> Acesso em: 29 jun. 2017.

QUINLAN, J. R. Induction of decision trees. **Machine Learning,** v. 1, p. 81-106, 1986.

RABACA, Carlos Alberto; BARBOSA, Gustavo. **Dicionário de Comunicação**. 2. ed. São Paulo: Ática, 1995. 637p.

RIBEIRO, Gerson et al. **Web Analytics:** uma visão brasileira II. 2012. Disponível em: <https://www.slideshare.net/gersonribeiro/ebook-web-analytics-uma-viso-brasileira-ii>. Acesso em: 21 mar. 2017.

ROZNER, Steve. **Developing key performance indicators:** a toolkit for health sector managers. Bethesda, MD: Health Finance & Governance Project, Abt Associates Inc., 2013. Disponível em: <https://www.hfgproject.org/wp-content/uploads/2014/10/03-Developing-Key-Performance-Indicators.pdf>. Acesso em: 27 mar. 2017.

SANTOS, N. **Gestão estratégica do conhecimento:** capítulo 1 – conhecimento organizacional. Apostila não publicada do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2005.

SELL, Denilson; CABRAL, Liliana; MOTTA, Enrico; DOMINGUE, John; PACHECO, Roberto. Adding semantics to business intelligence. In: **International Workshop on Web Semantics (WebS 2005) / 16th International Conference on Database and Expert Systems Applications (DEXA 2005)**, 22-26 Aug 2005, Copenhagen, Denmark. Disponível em: <http://oro.open.ac.uk/23148/>. Acesso em: 8 abr. 2017.

SERRA, J. Paulo. **Manual de teoria da comunicação**. Covilhã: Livros Labcom, 2007.

SVEIBY, K. E. **A nova riqueza das organizações**. Rio de Janeiro: Campus, 1998.

TUNDALWAR, Rashmi R.; KULKARNI, Manasi. Web spam detection using improved decision tree classification method. **International Journal of Computer Science & Information Technologies**, v. 5, n. 4, p. 4936-4942, jul. 2014.

UNIVERSIDADE DE CHICAGO. **Semistructured data model.** 2017. Disponível em: <https://ochre.uchicago.edu/page/semistructured-data-model>. Acesso em: 29 jun. 2017.

WAZLAWICK, Raul Sidney. **Metodologia de pesquisa para Ciência da Computação.** Rio de Janeiro: Elsevier, 2008.

W3TECHS. **Usage of traffic analysis tools for websites.** Disponível em: <https://w3techs.com/technologies/overview/traffic_analysis/all>. Acesso em: 19 mar. 2017.

ZENG, Li et al. Techniques, process, and enterprise solutions of business intelligence: systems, man and cybernetics. 2006. SMC'06. In: **IEEE International Conference on Systems, Man, and Cybernetics,** October 8-11. Taipei, Taiwan, v. 6, p. 4722, 2006.

ZOTOS, Dimitri. **20 KPIs you should monitor in Google Analytics.** 2011. Disponível em: <http://www.webseoanalytics.com/blog/20-kpis-you-should-monitor-in-google-analytics/>. Acesso em: 2 abr. 2017.

Árvore de Decisão para Web Analytics: uma Proposta de Diretrizes para Classificação de Métricas do Google Analytics

Marco A. R. Baumann

Departamento de Informática e Estatística – Universidade Federal Santa Catarina (UFSC)
Caixa Postal 476 – 88060-900 – Campus Universitário Trindade – Florianópolis – SC – Brasil

marco.baumann@grad.ufsc.br

Abstract. *It is intended with this work to enable the creation of a decision tree using data from the monitoring done with Google Analytics. For this, a step-by-step guide to collecting and transforming the metrics and dimensions collected by Google Analytics, its classification through the use of data mining techniques, is suggested, culminating in the creation of the decision tree to be used by the business managers. The suggested process seeks to use tools that do not reflect financial investment. As a result, it is hoped to encourage the use of metrics by managers of small and medium digital businesses, providing them with knowledge that enables greater success in decision making.*

Resumo. *É pretendido com este trabalho possibilitar a criação de uma árvore de decisão utilizando dados provenientes do monitoramento feito com Google Analytics. Para isto é sugerida uma diretriz passo-a-passo para coleta e transformação das métricas e dimensões coletadas pelo Google Analytics, sua classificação através do uso de técnicas de data mining, culminando na criação da árvore de decisão a ser utilizada pelos gestores de negócios. O processo sugerido busca utilizar ferramentas que não reflitam em investimento financeiro. Como resultado espera-se incentivar o uso de métricas por gestores de pequenos e médios negócios digitais, fornecendo-lhes conhecimento que possibilite maior sucesso em tomadas de decisão.*

1. Introdução

No mercado contemporâneo, a competitividade é uma das principais características para garantir a continuidade das organizações e seus negócios. Na da informação, a arma mais importante para a manutenção desta continuidade é, como o próprio nome diz, a informação: sobre o mercado, produtos, concorrentes, e a mais preciosa delas: sobre o cliente. É utilizando-se deste tipo de conhecimento que o negócio pode não só se manter, mas também se renovar, atendendo melhor a necessidade de seus clientes. Sendo assim imprescindíveis estes elementos na tomada de decisão de inovação que poderá resultar na fidelização seu público, evitando a perda deste para os seus concorrentes.

Como resposta à esta nova realidade de mercado ocorre a popularização do uso de ferramentas de avaliação de desempenho de ações e métricas (FERNANDES; ROSA, 2013, p. 183), antes somente disponíveis para grandes empresas mediante altos investimentos. Tais ferramentas tornaram-se imprescindíveis para aferir e garantir a manutenção de negócios atuantes na internet.

Os outputs dessas ferramentas possibilitam a tomada de decisão com base em dados estatísticos, de forma mais científica, em detrimento de tomadas de decisão puramente intuitivas. Estes outputs servirão de base para a concepção de novas oportunidades, ou até mesmo para mudanças de rotas na gestão estratégica dos negócios (LISBOA, 2012).

A presente pesquisa tem como foco o fornecimento de conteúdo relevante e necessário para a elaboração de um plano de monitoramento eficaz, evitando a ocorrência da sobrecarga cognitiva ou desinformação. Sua realização iniciará pelo levantamento das principais ferramentas de web analytics disponíveis no mercado, categorização de tipos de negócios digitais, avaliação das principais métricas e KPIs fornecidos pela ferramenta escolhida, catalogação dos dados gerados, suas classificações e relevância relativa ao contexto, resultando na criação de uma metodologia disponível aos gestores para que as situações de sobrecarga cognitiva sejam evitadas.

1.1 Solução Proposta

Como solução para o problema apresentado, o objetivo de pesquisa é a proposta de uma diretriz para a construção de uma árvore de decisão utilizando softwares gratuitos, tendo como base dados fornecidos pela ferramenta de web analytics Google Analytics em sua versão gratuita, possibilitando aos gestores de negócios digitais uma alternativa mais simples para a aquisição de conhecimento sobre seus negócios digitais por meio de BI e data mining sem a necessidade de investimentos financeiros e consequentemente tornando conhecimento mais acessível ao mercado e às pequenas e médias empresas.

2. Metodologia da Pesquisa

Seguindo os preceitos de Wazlawick (2008), para a obtenção dos objetivos, tanto gerais quanto específicos, deste trabalho foram realizadas pesquisas bibliográficas sobre os temas abordados, desde os fundamentos da gestão do conhecimento, conceitos de dados, informação e conhecimento, sobre a natureza de web analytics modernos, tipologia de websites e suas principais métricas e indicadores de desempenho.

Sob a ótica de sua natureza, este trabalho tem com algo a geração de conhecimento para a aplicação prática de solução para um problema específico, classificando-se assim o trabalho como uma pesquisa aplicada.

Sob o ponto de vista de seus objetivos, esta pesquisa é exploratória, pois envolver um levantamento bibliográfico, análise de conceitos e classificações de autores e validação da solução proposta..

3. *Key-Performance Indicators*

KPIs também têm uma relação e são utilizadas como base para avaliações tanto quantitativas quanto qualitativas do negócio. Farris (2013) define métricas como o sistema de mensuração que quantifica uma tendência uma dinâmica ou característica. Sendo estas utilizadas na explicação de fenômenos, identificar causas, criando a possibilidade de comparação de tais observações em diferentes espaços de tempo. Portanto são somente as medidas geradas pela avaliação de um acontecimento, não tendo necessariamente relação com os objetivos da organização, sendo este seu principal ponto de diferenciação para com os indicadores-chave de performance, como afirmado por Rozner (2013) em seu relatório para a Agência dos Estados Unidos Para O Desenvolvimento Internacional (USAID), a definição mais simples que se pode ter de um KPI é: medidas que um setor ou organização utiliza para definir seu sucesso e avaliar o progresso na conquista de seus objetivos estratégicos.

No contexto desta pesquisa, o valor semântico relativo às informações representadas pelos KPIs é uma peça fundamental para a classificação destes nos os possíveis conjuntos de análise. Este valor semântico pode variar de acordo com o tipo de negócio ou website analisado, existindo um grupo de KPI que poderiam ser considerados principais, que são de interesse para qualquer tipo de website, e outros que pode ou não ter valor, dependendo do mercado.

Por este motivo, é importante que na etapa de seleção e criação de indicadores, deve-se iniciar pelo principal objetivo a ser alcançado, para então a partir deste elaborar os objetivos secundários e finalmente, criar os *KPIs* referentes (POPA, 2015).

4. Solução Proposta

A solução proposta é coleta as métricas selecionadas através da API de desenvolvedores do Google Analytics, isto é feito utilizando criando uma conexão de serviço entre do Pentaho PDI, onde, após coletados, são pré-processados para que respeitem as características requeridas para a sua futura classificação. Os dados são então salvos no sistema de arquivo em formato ARFF – Attribute-Relation File Format – no qual é lido pelo programa de data mining Weka para sua classificação utilizando-se o algoritmo J48 para elaboração da árvore de decisão, que é o produto final do processo. Todas as macro-etapas do processo estão descritas na Figura 1.

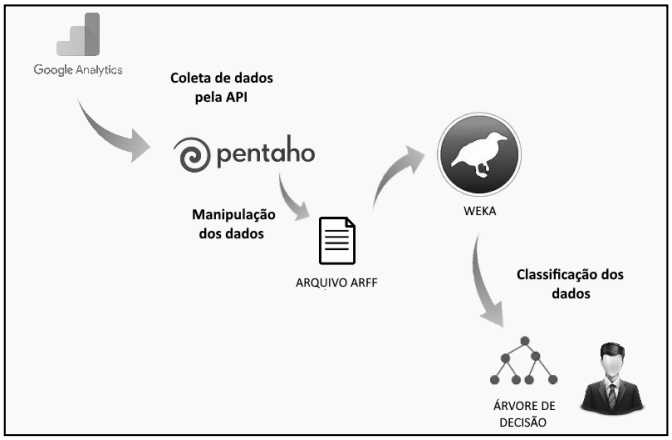


Figura 1. Visão geral do processo sugerido

4.1. Visão Geral

Com objetivo de definir quais são as regras de negócio e requisitos de uso para execução da solução proposta, serão listadas a seguir as ferramentas utilizadas, definindo seu propósito no processo.

4.1.1 Coleta de Dados pela API Google Analytics

Fase inicial do processo, onde o software Pentaho conecta diretamente à API do Google Analytics, são definidas as configurações da propriedade em questão e escolhidas as métricas desejadas para a classificação, o período de avaliação e a granularidade dos dados para coleta das métricas solicitadas. Esta solicitação deve ser feita utilizando-se de uma conta de desenvolvedor devidamente habilitada e com identidade verificada no momento da conexão, isto sendo descrito nos requisitos funcionais desta etapa.

4.1.2 Manipulação dos dados Pentaho

Logo após a coleta de dados, este são filtrados para remoção de dados complementares enviados pela API juntamente aos dados solicitados, pois não tem valor para a classificação posterior. Nesta etapa também são alteradas as tipificações de dados, enquadrando-os nos requisitos da etapa de classificação. Finalizando com a criação do arquivo ARFF, salvo no sistema de arquivos do sistema operacional.

4.1.2 Classificação dos dados Weka

Após o carregamento do arquivo ARFF gerado pelo Pentaho na etapa anterior, no Weka é feita a preparação dos dados para sua classificação e configuração das opções do algoritmo J48, equivalente Weka do algoritmo de classificação C4.5, também são feitas as seleções de métricas a fim de aumentar a qualidade do resultado da classificação, resultando por fim, a árvore de decisão referente às métricas e a conversão escolhidas.

5. Experimentos Realizados

Para a avaliação da diretriz, esta foi aplicada em dois websites, dos quais foram selecionadas treze métricas do Google Analytics, sendo uma delas o atributo classe "ga:goalCompletionsAll". As restantes são: mídia de acesso (ga:medium), hora de acesso (ga:hour), tipo do dispositivo utilizado para acesso (ga:deviceCategory), categoria de usuário (ga:userType), dia da semana do acesso (ga:dayOfWeek), dias desde a última visita (ga:daysSinceLastSession), número de visitas (ga:sessions), rejeições (ga:bounces), visualizações de página (ga:pageViews), tempo na página (ga:timeOnPage), duração da visita (ga:sessionDuration) e interações (ga:hits).

5.1 Experimento Website 01

Para avaliar o website da escola, foi selecionado o período de 30 dias anteriores à data de realização do teste, que ao serem agregados por minuto, resultou em 9.225 registros. Dentre estes registros, 4.373 registros foram identificados como ocorrência de evento/conversão, 47,4% do total.

O número de registros classificados de forma correta no experimento foi de 9.188, representando uma Acurácia de 99,59%, já a precisão alcançada foi de 99,72%, com 4.361 positivos-verdadeiros e 12 falsos-negativos. Outras características estão descritas no Quadro 6, que apresenta o output bruto do Weka. Sendo convertido o resultado para uma representação gráfica (Figura 2), apresenta-se a seguinte árvore de decisão.

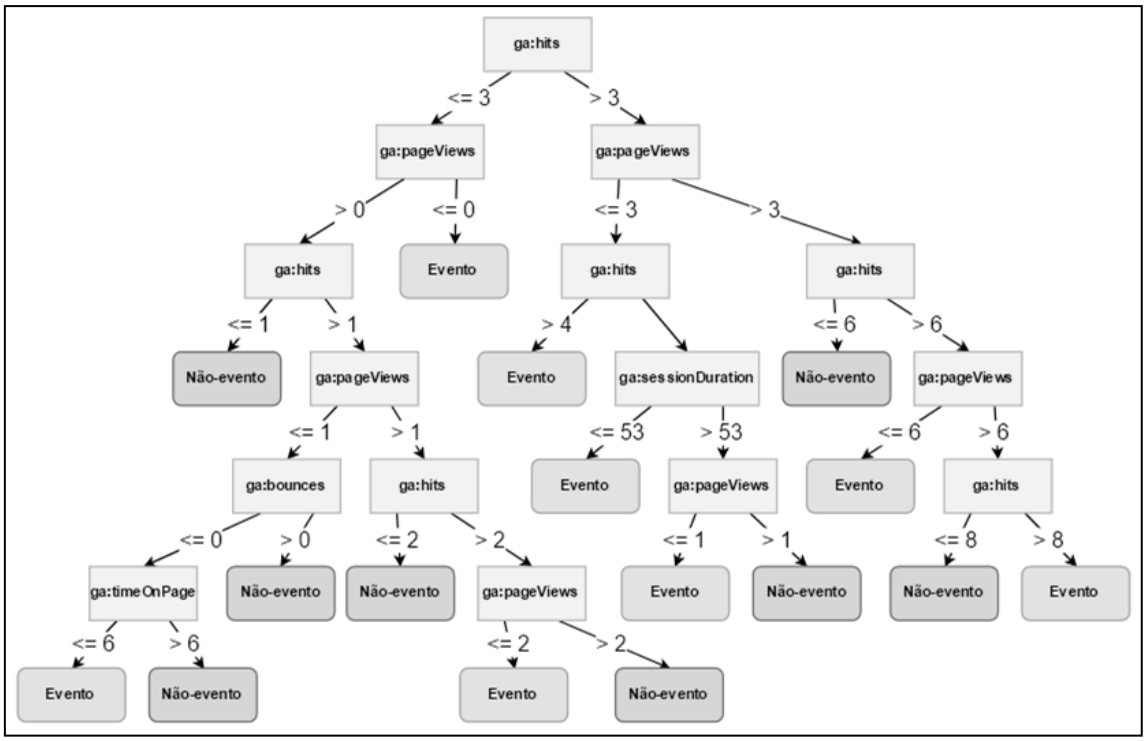


Figura 2. Árvore de Decisão Website 01

Em sua forma podada, resultante do algoritmo J48, esta árvore tem tamanho – quantidade de nodos e folhas – 31 e 16 folhas, descrevendo a ocorrência ou não-ocorrência do evento analisado. Não apresenta uma legibilidade tão facilitada quanto esperado, porém é possível compreender a dinâmica entre as métricas e sua influência no resultado final, na classe-folha da árvore.

5.2 Experimento Website 02

Para avaliar o website 02, foi selecionado o período de em que se iniciou o monitoramento de eventos de 10 dias de dados que, ao serem agregados por minuto, resultou em 5503 registros. Dentre estes registros, 68 registros foram identificados como ocorrência de evento/conversão, 1,02% do total.

O número de registros classificados de forma correta neste experimento foi de 5.482, representando uma acurácia de 99,61%, ponderada, já a precisão alcançada foi de 79,41%, com 54 positivos-verdadeiros e 14 falsos-negativos. Outras características estão descritas no Quadro 7, que apresenta o output bruto do Weka. Sendo convertido o resultado para melhorar sua legibilidade a classificação resultante é representada pela seguinte árvore de decisão (Figura 03).

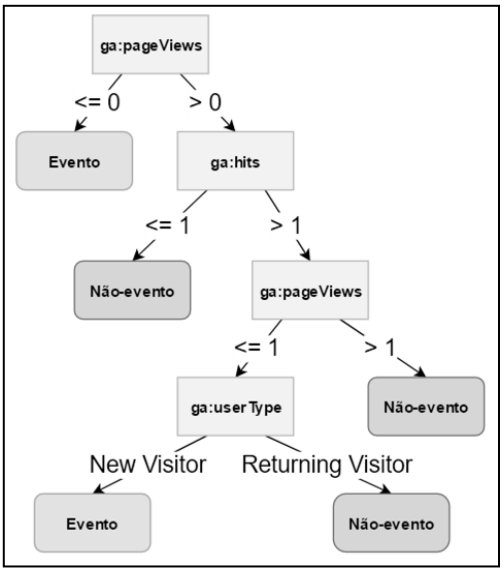


Figura 3. Árvore de decisão Website 02

A árvore de menor tamanho, 9 e somente 5 folhas, sendo consideravelmente menor que a árvore do experimento número 01, possivelmente influenciado pela quantidade inferior de registros de eventos existentes na base de dados, mas com alta legibilidade.

6. Conclusão

Percebe-se com estes dois experimentos, dois resultados finais distintos, sendo no experimento do Website 01, que a árvore apresenta um tamanho próximo do esperado pelo autor. No experimento Website 02, temos a aparição de uma variável binária, mostrando que estas podem contribuir ativamente para a ocorrência da classe evento. Porém, por mais que as árvores de decisão apresentam indicadores de precisão e acurácia aceitáveis, ao ser feita uma análise mais minuciosa quanto aos valores de atributos selecionados como nodos da árvore, como por exemplo na Figura 04, vemos situações que não condizem com uma situação real, onde com número de visualização de página (pageviews) menor ou igual a zero, resulta-se em evento.

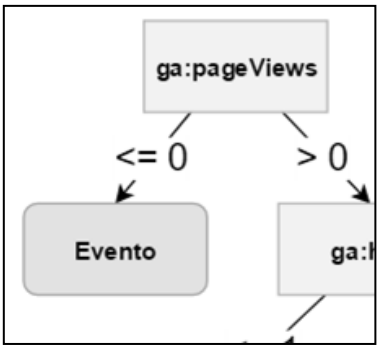


Figura 4. Detalhe da árvore de decisão referente ao Website 02

Avaliando se esta característica advém dos dados coletados ou se é uma anomalia resultante do processo de coleta e transformação dos dados, percebeu-se que a informação consta já na base de dados do Google Analytics. Este fato levantou algumas dúvidas:

- Seria o Google Analytics uma ferramenta eficaz para esta etapa do processo? Quais suas limitações?
- Há alguma característica dos websites avaliados que resultam nesta situação? Qual o impacto que a estrutura do website pode ter nas métricas que o Google Analytics coleta?

Para a primeira dúvida, será necessário aprofundar o estudo para elucidar a forma de obtenção destas métricas tidas como anômalas, para assim entender o que pode causa-las com mais precisão. Sendo o Google Analytics uma ferramenta utilizada intensivamente pelo mercado mundial para medição de desempenho de websites, pode-se descartar a possibilidade que seja algum problema estrutural na aplicação, o que nos leva à segunda dúvida, dando foco nos objetos monitorados pelo Google Analytics, terão estes websites alguma característica que causa tais anomalias nas métricas? No caso do Website 02, sabe-se que o evento avaliado ocorre na página principal do website, já no Website 01, este evento só é realizado em uma subpágina do website. Junto a isto temos o volume de eventos registrados, assim como a seleção de métricas a serem utilizadas na criação da árvore de decisão, estes podem ser fatores determinantes na qualidade do conhecimento gerado ao fim do processo. Para chegar a uma conclusão assertiva sobre isto será necessário um estudo específico para obtenção de resultados, ficando estas sugestões para trabalho posteriores nesta linha de estudo.

7. Trabalhos Futuros

O método proposto neste trabalho se prova eficaz para a criação do artefato de conhecimento pretendido para auxiliar gestores iniciantes na atuação de data mining e BI para seus negócios, porém sua versão apresentada podem receber melhorias para potencializar sua eficiência e qualidade do conhecimento gerado. Entre as melhorias possíveis, as listadas a seguir podem ser de grande valia:

- Ajustes das variáveis de configuração do J48: dependendo do website avaliado, suas métricas e suas características poderão responder melhor às configurações diferentes do J48, possibilitando assim mais qualidade no conhecimento gerado;
- Testar outras métricas e dimensões e avaliar AD resultante: a ferramenta Google Analytics disponibiliza mais de 50 métricas e dimensões sobre o website/aplicação analisado, buscar formas de coletar todos estes dados respeitando os limites da API dará mais opções de avaliação por parte do gestor e também ajuda a entender o impacto na qualidade da árvore gerada;

• Testar a qualidade do conhecimento gerado: entender como os gestores podem utilizar este conhecimento e adequá-lo ao contexto de uso, tornando-o mais útil e relevante aos gestores;

• Como o Google Analytics coleta as métricas utilizadas: saber exatamente como os dados são gerados pode trazer melhor conhecimento sobre características do website e como métricas de baixa qualidade podem ser evitadas;

• Avaliar a relação entre estrutura do website e os dados coletados pelo Google Analytics: para compreender como a arquitetura de informação do website pode influenciar as métricas coletadas pelo web analytic e consequentemente impacta na árvore de decisão gerada.

Diversos ajustes e operações de transformação de dados podem ser aplicadas durante a etapa de manipulação dos dados no Pentaho, estando isto a critério e criatividade do profissional que realizar o método proposto neste trabalho. A análise de informações de negócios deve sempre trazer perspectivas promissoras aos seus gestores e este é apenas o primeiro passo em direção a concretização deste objetivo.

8 Referências

Farris, Paul W. et al. Métricas de marketing: o guia definitivo de avaliação de desempenho do marketing. 2. ed. Porto Alegre: Bookman, 2013. 426p.

Fernandes, Bruno G.; Rosa, Celso O. As métricas do marketing no século XXI. Revista Panorama, Goiás, v. 3, n. 1, p. 180, jan./dez. 2013. Disponível em: <http://estudos.ucg.br/index.php/panorama/article/download/3435/2006>. Acesso em: 22 jun. 2015.

Lisboa, Ruben. Importância do Web Analytics para o seu negócio online. 2012. Disponível em: <https://marketingdigitalpt.wordpress.com/2012/09/03/importancia-web-analytics-negocio-online/>. Acesso em: 9 out. 2015.

Popa, Brînduşa Maria. *Challenges when developing performance indicators. Journal of Defense Resources Management (JoDRM)*, Brasov, Romania, n. 1, ano 6, p. 111-114, out. 2015. Disponível em: <https://www.ceeol.com/search/article-detail?id=305960>. Acesso em: 8 abr. 2017.

Rozner, Steve. *Developing key performance indicators: a toolkit for health sector managers*. Bethesda, *MD*: *Health Finance & Governance Project*, *Abt Associates Inc*., 2013. Disponível em: <https://www.hfgproject.org/wp-content/uploads/2014/10/03-Developing-Key-Performance-Indicators.pdf>. Acesso em: 27 mar. 2017.

Wazlawick, Raul Sidney. Metodologia de pesquisa para Ciência da Computação. Rio de Janeiro: Elsevier, 2008.