https://doi.org/10.47456/bjpe.v7i3.35743

NÚCLEO DE PESQUISA EM GESTÃO DE SISTEMAS DA PRODUÇÃO - NP GSP

Brazilian Journal of Production Engineering
BJPE - Revista Brasileira de Engenharia de Produção
Disponível em: http://periodicos.ufes.br/BJPE/index

ISSN: 2447-5580

Campus São Mateus
UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

ARTIGO ORIGINAL OPEN ACCESS

APLICAÇÃO DE UM MODELO DE DESCOBERTA DE CONHECIMENTO NA ERA DO BIG DATA

APPLICATION OF A KNOWLEDGE DISCOVERY MODEL IN THE ERA OF BIG DATA

Emerson Rabelo [1*], Fernando Celso de Campos [2], & Leandro Magno Correa da Silva [3]

[1 3] IFPR – Instituto Federal do Paraná – Campus Avançado de Astorga. [2] UNIMEP – Universidade Metodista de Piracicaba - PPGEP/FEAU

[1*] emerson.rabelo@ifpr.edu.br [2] fccampos@unimep.br [3] leandro.silva@ifpr.edu.br

ARTIGO INFO.

Recebido em: 14.06.2021

Aprovado em: 27.07.2021

Disponibilizado em: 16.08.2021

PALAVRAS-CHAVE:
Desenvolvimento de Produto; Descoberta de Conhecimento; Big Data; Rede Social.

KEYWORDS:
Product Development; Knowledge Discovery; Big Data; Social Network.

***Autor Correspondente**: Rabelo, E.

RESUMO

A facilidade e a evolução do acesso tecnológico têm sido responsáveis pela velocidade e pelo volume com que os dados são produzidos. Em consequência, surgem cenários, oportunidades e desafios que favorecem as tomadas de decisão e auxiliam o processo de desenvolvimento do produto (PDP). Diante disso, o presente trabalho desenvolveu o modelo denominado MDC-PDP (Modelo de descoberta de conhecimento no processo de desenvolvimento de produto), com objetivo de apoiar a descoberta de conhecimento no processo de desenvolvimento do produto. Para dar suporte ao modelo, foram utilizadas as metodologias tradicionais associadas às demandas do Big Data. Para ilustrar a sua aplicação, o presente modelo foi aplicado no domínio de aplicação da indústria de moda. O modelo evidenciou que esforços empreendidos na compreensão antecipada dos dados, podem contribuir para que os dados extraídos sejam menos complexos. Outra evidência é a dissociação entre volume e valor dos dados, pois o valor dos dados não está vinculado ao seu volume. Por fim o MDC-PDP também contribuiu na obtenção de conhecimentos úteis e novos no desenvolvimento de coleção de moda, sendo possível aplicar o modelo em outros domínios de aplicação.

ABSTRACT

The ease and evolution of technological access has been responsible for the speed and volume with which data is produced. As a result, scenarios, opportunities and challenges arise that favor decision-making and help the product development process (PDP). Therefore, the present work developed the model called MDC-PDP (Knowledge Discovery Model in the Product Development Process), aiming to support the knowledge discovery in the product development process. To support the model, traditional methodologies associated with Big Data demands were used. To illustrate its application, this model was applied in the application domain of the fashion industry. The model showed that efforts made in the early understanding of the data, can contribute to the extracted data being less comple. Another evidence is the dissociation between volume and data value, as the data value is not tied to its volume. Finally, the MDC-PDP also contributed to obtaining useful and new knowledge in the development of fashion collections, making it possible to apply the model in other application domains.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

1. INTRODUÇÃO

Atualmente, em virtude do crescente aumento no volume de informações vem se instalando um novo cenário em que surgem novos desafios aliados a novas necessidades de consumidores cada vez mais exigentes. As oportunidades geradas por esses novos desafios podem favorecer as tomadas de decisão e auxiliar no processo de desenvolvimento de produto - PDP. A competitividade e a velocidade dos processos nas empresas exigem que as tomadas de decisão e o desenvolvimento de estratégias sejam realizados com base em conhecimentos úteis e concretos. Assim, os gestores adquirem diferentes visões das mais variadas dimensões na dinâmica da empresa, passando a se interessar pela criação de um diferencial na relação entre empresa/cliente, o que cria a possibilidade de surgirem novas ideias para o desenvolvimento de produtos.

A descoberta de conhecimento por meio de dados estruturados e internos, que já era um possível suporte à decisão, se faz agora por meio de dados não estruturados ou semiestruturados, intitulado Big Data. Em vista disso, o *Big Data* pode oferecer um diferencial competitivo estratégico, especialmente quanto a: *i)* identificar em tempo real os anseios dos consumidores; *ii)* evidenciar os defeitos dos produtos; *iii)* gerar oportunidades para a otimização dos produtos e *iv)* apontar tendências e hábitos de consumo.

Com o advento do *Big Data*, surge a oportunidade de se obterem novas visões e dimensões em relação aos processos e produtos da organização e, assim, de se revolucionar a tomada de decisão durante a fabricação e a venda desses produtos.

A utilização de diversas ferramentas tecnológicas por consumidores e empresas, gera volume significativo de dados relacionado a produtos, outras características dos dados geradas por essas ferramentas, estão na velocidade com que são produzidos e os diferentes formatos estruturais. Com base nessas características (volume, velocidade e variedade) o objetivo do presente trabalho é apresentar um modelo de descoberta de conhecimento que auxilie a tomada de decisão no processo de desenvolvimento do produto, sendo assim, com interesse de demonstrar sua efetividade, o modelo proposto, é utilizado no domínio de aplicação relacionado à indústria da moda.

2. DESCOBERTA DE CONHECIMENTO

Fayyad, Piatetsky-Shapiro e Smyth (1996) dissertaram sobre o empolgante ritmo na coleta e no acúmulo de dados e alertaram para a urgente necessidade de uma nova geração computacional e de ferramentas tecnológicas que auxiliem na extração de conhecimentos do crescente volume de dados digitais. Tal preocupação permanece atual, pois, embora os avanços tecnológicos tenham apoiado as descobertas de conhecimento, estas geraram novas oportunidades e criaram novos desafios.

A descoberta de conhecimento em banco de dados (*Knowledge Discovery in Database* ou KDD) foi formalizado em 1989 para nomear a abordagem destinada a atender aos processos referentes à busca de conhecimento a partir de bases de dados. Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 30) propuseram a definição que se tornou a mais popular na literatura: "KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados".

Duas décadas depois, a análise dessa definição permite que o mesmo propósito seja mantido quando se trata de associá-la às novas demandas do *Big Data*. Hashem *et al.* (2015) argumentam que a natureza do *Big Data* é indistinta e envolve consideráveis processos para identificar e converter os dados em novas ideias.

Além do processo de descoberta de conhecimento em banco de dados (KDD), existe também o modelo Crisp-DM, o qual estabelece um conjunto de regras e tarefas para a orientação do processo de descoberta de conhecimento. O CRISP-DM é uma metodologia não proprietária que foi criada em 1996, por um consórcio de empresas e consumidores que atuam na área de mineração de dados.

Essas exigências são criadas pela produção e consumo de um volume significativo de dados, os quais, com velocidade cada vez maior, são gerados por diversas ferramentas e adquirem diferentes formatos estruturais.

Mcafee e Brynjolfsson (2012) afirmam que, por dia, ocorre um aumento de 2,5 *exabytes* na produção de dados; Davenport (2014), por sua vez, ressalva que o mundo utilizou 2,8 *zetabytes* de dados em 2012, mas que apenas 0,5% desses dados foram analisados de alguma forma. Esse autor estima que aproximadamente 25% deles têm valor potencial e reconhece que essa estimativa é modesta quando se considera a quantidade de dados disponíveis. Essa evidência incentiva a realização de pesquisas sobre a área de MD, inclusive algoritmos de dados quantitativos.

As definições de *Big Data* na literatura convergem quanto aos seguintes fatos: utilização de diferentes fontes de dados e características como tipo de dados, volume, velocidade e variedade (Manyika *et al.*, 2011; Begoli & Horey, 2012; Mcfee & Brynjolfsson, 2012; Kaisler, Armour, Espinosa, & Money, 2013; Davenport, 2014; LI, Tao, Cheng & Zhao, 2015; Gantz, Reinsel, & Shadows, 2012). Estendendo a definição, Zikopoulos *et al.* (2011) acrescenta a característica veracidade e Kaisler *et al.* (2013) mencionam as características valor e complexidade. Davenport (2014) agrega venalidade, isto é, a possibilidade de ser vendido.

3. MODELO MDC-PDP

O modelo de descoberta de conhecimento para o processo de desenvolvimento de produto (MDC-PDP) desenvolvido nesta pesquisa tem como base a metodologia Crisp-DM, a qual não está sendo atualizada para as demandas do *Big Data* e tampouco para a ciência de dados moderna (Piatetsky, 2014; Asamoah & Sharda, 2015).

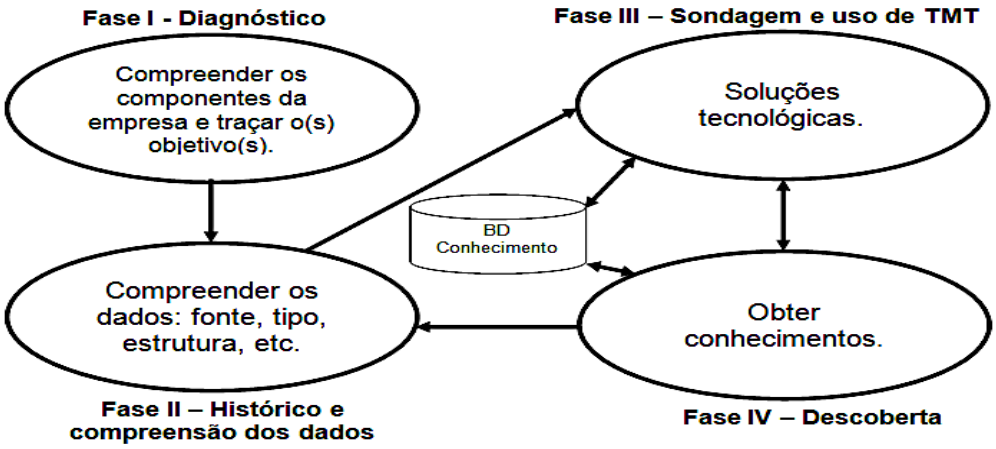
Os autores Schröer e Gómez (2021) realizam uma revisão sistemática da literatura sobre a aplicação do modelo de processo CRISP-DM, e concluem que a necessidade de potenciais melhorias no CRISP-DM devem ser revistas, pois identificaram que falta uma fase de implantação. Também consideraram que pesquisas futuras devem explorar maneiras adequadas de integrar modelos em um ambiente produtivo.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering*, *7*(3), 106-125.

O MDC-PDP tem características genéricas, o que torna possível realizar o processo de descoberta de conhecimento do início ao fim do processo de descoberta de conhecimento, e atender às demandas do *Big Data* sem deixar de considerar os métodos tradicionais de MD.

O modelo MDC-PDP é composto por quatro fases, como mostra a Figura 1, nas quais se buscam organizar as atividades de descoberta de conhecimento, além de auxiliar a compreensão das soluções tecnológicas necessárias.

Figura 1. Fases do MDC-PDP



Fonte: Autores

A primeira fase, “Diagnóstico”, corresponde às atividades que permitem compreender a empresa, isto é, obter familiaridade no processo de negócio e no domínio de aplicação. A segunda, “Histórico e compreensão dos dados”, refere-se ao levantamento das fontes de dados e suas características, cujo foco é a qualidade da informação, esta fase compreende quatro etapas: *i)* identificação das fontes de dados; *ii)* facetas para a fonte de dados *iii)* avaliação das características do conjunto de dados *iv)* e decisão das soluções tecnológicas. A terceira, “Sondagem e uso de TMT (Técnicas, Métodos e Tecnologias)”, é composta pelo levantamento de técnicas, métodos e tecnologias utilizadas atualmente, implementada pelas etapas: i) preparação dos dados e ii) solução de armazenamento. A quarta, “Descoberta”, caracteriza-se pela etapa de: *i)* análise dos dados; e, ii) técnicas de visualização.

3.1 APLICAÇÃO DO MDC-PDP

O MDC-PDP foi empregado em uma empresa de confecção industrial, em atividade há aproximadamente oito anos, com matriz situada na cidade de Maringá, no estado do Paraná. Para tomar decisões quanto ao desenvolvimento de novos produtos, a referida indústria utiliza informações relacionadas às tendências de moda feminina, como cores, estilos, temas, marcas e designer, e também a personalidades em destaque nas mídias

O desenvolvimento dos produtos ocorre sazonalmente, sendo a produção dos diferentes períodos denominada de “coleção”. Em reuniões realizadas com os colaboradores da indústria, os mesmos relataram que alguns dos indicadores de tendências utilizados como parâmetro para a produção dessas coleções são os desfiles internacionais de moda. De acordo com o diretor, os eventos internacionais influenciam as tendências do ano seguinte no Brasil.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

Segundo diretor da indústria de confecção, decisões não assertivas resultaram em prejuízos. Caso fossem obtidas mais informações sobre as tendências e os temas da moda, algumas dessas decisões poderiam ter sido mais assertivas e menos prejudiciais.

Os resultados da aplicação do MDC-PDP, além de terem sido avaliados pelos colaboradores da referida indústria, também foram avaliados por outros colaboradores de uma grande indústria que atende boa parte do Brasil.

3.1.1 Fase I: Diagnóstico

A Fase I do modelo MDC-PDP tem como objetivo compreender os componentes da empresa e traçar o(s) objetivo(s).

A indústria de confecção dispõe de uma equipe de estilistas que, apoiada pelo setor de marketing da indústria, realiza levantamentos e coletas dos elementos referentes aos temas e tendências da moda. Para tanto, a equipe acompanha os eventos de moda, as revistas, os portais eletrônicos, os comentários e as opiniões de especialistas e até mesmo as influências sociais.

Tais influências sociais têm se destacado ultimamente em razão do surgimento de pessoas que, nos canais de comunicação disponíveis na internet, partilham opiniões, ideias, conceitos e críticas com seus seguidores. Para Halvorsen, Hoffmann, Coste-Manière e Stankeviciute (2013), os comentários de moda produzidos e divulgados nesses canais influenciam o comportamento de compra e também estimulam o consumo dos seus seguidores. Assim, os estilistas e o departamento de marketing, desenvolvem trabalhos de pesquisa para identificar esses influenciadores e seus comportamentos no mercado da moda.

De acordo com os colaboradores da indústria, os eventos de moda internacional, especialmente os ocorridos na Europa e nos EUA, repercutem no Brasil com atraso médio de um ano. Com essa informação, foi elaborada uma lista com alguns pontos que podem pautar a definição do objetivo desta fase: *i)* produtos concorrentes e similares; *ii)* novos nichos de clientes em potencial; *iii)* necessidades dos clientes; *iv)* requisitos dos clientes sobre o produto; *v)* requisitos do produto; *vi)* ideias para novos produtos.

3.1.2 Fase II - Histórico e Compreensão dos Dados

Esta Fase tem como objetivo conhecer os dados já produzidos no domínio de aplicação, ou seja, os dados que foram explicitados e registrados. Seguindo o objetivo traçado na primeira fase, é possível manter o foco no que se deseja e, dessa forma, pesquisar e identificar as fontes interna e externa de dados.

Etapa 1 – Identificação da fonte de dados

Para identificação das possíveis fontes de dados, foram realizadas reuniões, nas quais os colaboradores expuseram os procedimentos realizados para o levantamento das informações que subsidiavam as tomadas de decisão durante o desenvolvimento da coleção. Os destaques incidiram sobre a obtenção de dados nas mídias sociais, cujos usuários produzem comentários e discussões sobre eventos de moda. Nesses ambientes, sem limite geográfico, diferentes usuários se conectam para compartilhar conteúdos e trocar experiências.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

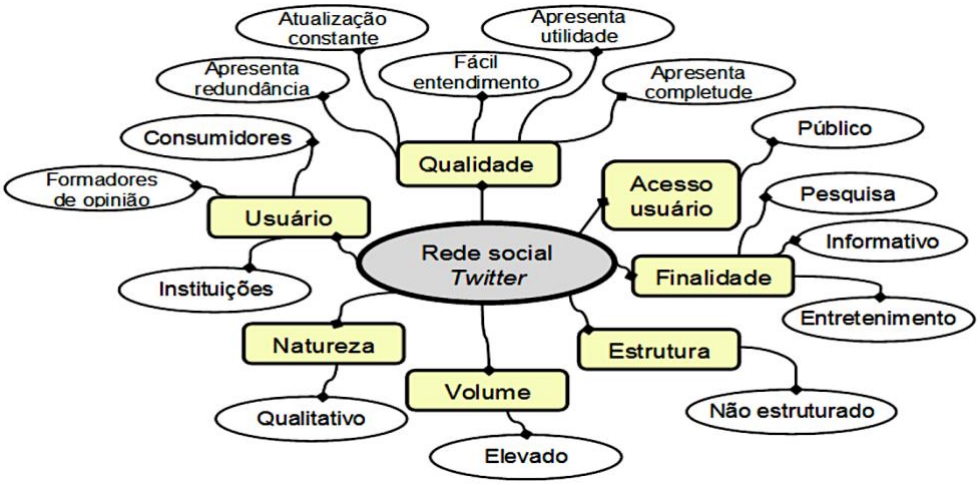
Nesse caso específico, tendo em vista sua objetividade e a possibilidade de vincular termos específicos a eventos de moda, dentre as mídias sociais disponíveis, optou-se pela utilização de conjuntos de dados originados da rede social *Twitter*.

Etapa 2 – Facetas para a fonte de dados

O MDC-PDP, utiliza a análise de faceta com o intuito de auxiliar a compreensão das fontes de dados. A análise de faceta é conhecida na área de ciência da informação e tem sido amplamente utilizado como mecanismo em vários sistemas de organização do conhecimento; dentre eles, os sistemas de classificação, taxonomias, incluindo o desenvolvimento de arquiteturas de sites e de estruturas de informação visual (Shiri, 2014). Em virtude dos benefícios da aplicação da análise de facetas, suas classificações têm sido amplamente exploradas na organização e na recuperação de informações no domínio da Web (Milonas, 2011).

Para melhor compreensão da fonte de dados, realiza-se o desenvolvimento de "faceta" e "subfacetas", as quais representam, de forma generalizada, as características da fonte de dados. A Figura 2 ilustra as subfacetas derivadas da rede social *twitter* (faceta).

Figura 2. Faceta e Subfacetas da Rede Social Twitter



Fonte: Autores

Etapa 3 – Avaliação das características do conjunto de dados

Os conjuntos de dados utilizados nesta aplicação foram extraídos da fonte originada do *twitter*, com auxílio da linguagem R, que também é útil na manipulação desses dados. Foram utilizados os termos que fazem referência aos eventos de moda New York Fashion Week (NYFW) que ocorreu entre os dias 09 e 17 de fevereiro de 2017 e Milan Fashion Week (MFW) realizado entre os dias 22 e 28 de fevereiro de 2017. Foram extraídas 9.478 postagens para o conjunto de dados NYFW e 20.225 postagens para o conjunto de dados MFW.

Após a extração e a realização de uma breve análise dos conjuntos de dados, foi possível obter informações para responder ao formulário de detalhamento apresentado no Quadro 1. Esse formulário é integrado ao MDC-PDP.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering*, *7*(3), 106-125.

QUADRO 1. FORMULÁRIO DE DOCUMENTAÇÃO E DETALHAMENTO DAS CARACTERÍSTICAS DO CONJUNTO DE DADOS.

Detalhamento do conjunto de dados				
1. Objetivo da descoberta de conhecimento				
Obter conhecimentos referentes a produtos por meio de comentários em mídias sociais de eventos de referência em moda.				
2. Informações do conjunto de dados				
2.1 - Fonte(s): Rede social - Twitter				
2.2 - Formato(s) (variedade): ( ) Estruturado (**x**) Não Estruturado ( ) Semiestruturado				
2.3 - Atualização (alimentação da fonte de dados): (**x**) Tempo real (**x**) Diário ( ) Semanal ( ) Mensal ( ) Anual ( ) Outras:_____				
2.4 - Intervalo de tempo - (período inicial e final): 29 de fevereiro a 07 de marco de 2017.				
2.5 - Gerador (autor da fonte): usuários da rede social				
2.6 - Observações: termos-chaves de busca (NYFW; NYFW17; MFW; MFW17; MILANFASHION).				
3. Acompanhamento das características do conjunto de dados				
Características	**Avaliação**			**Observações**
	1ª	**2ª**	**3ª**	
3.1 - Credibilidade	**3**	**3**		
3.2 - Veracidade	**3**	**3**		
3.3 - Imparcialidade	**2**	**2**		
3.4 - Utilidade	**4**	**4**		
3.5 - Atualidade	**4**	**4**		
3.6 - Complexidade	**4**	**3**		
3.7 - Objetividade	**5**	**5**		
3.8 – Inconsistência	**4**	**3**		
3.9 - Apresenta qualidade mínima?	**Sim ( )**	**Sim (X)**	**Sim ( )**	
	Não (X)	**Não ( )**	**Não ( )**	
Legenda: (1) muito baixa- (2) baixa - (3) média - (4) alta - (5) muito alta				
4. Lista de recursos tecnológicos				
4.1 - Hardware: Computadores de baixo porte				
4.2 - Disponibilidade de armazenamento: Em torno de 10 Terabytes				

Fonte: Autores

No item 3 do formulário, que acompanha as características do conjunto de dados foram avaliadas algumas características do conjunto de dados. As avaliações assumem pontuações que variam de 1 a 5, sendo que 1 representa intensidade muito baixa e 5, muito alta. Como a análise é individualizada, a intensidade atribuída às características assume significado distinto, isto é, a relevância da intensidade independe da pontuação designada para as características. Por exemplo, o ideal é que a intensidade da característica inconsistência seja próxima do nível 1 e a da credibilidade, próxima do nível 5.

Por convenção, a "qualidade mínima", apresentada no item 3.9 do formulário, indica se todas as características possuem pontuação próxima do nível de intensidade desejado. Com base nisso, pode-se considerar o conjunto de dados habilitado para o processo de descoberta de conhecimento.

Observa-se que os itens 3.6 e 3.8 do Quadro 1, sofreram variação da primeira para a segunda avaliação, por conta dos tratamentos realizado na base de dados, que possibilitou a qualidade mínima para o processo de descoberta de conhecimento.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

A avaliação inicial do conjunto de dados mantém sua natureza bruta; no entanto, ao longo das fases do modelo proposto, são executadas atividades que alteram esse conjunto, o que implica a necessidade de uma nova avaliação das características. As características avaliadas foram:

- *Credibilidade* - Os geradores da fonte possuem credibilidade? Sabe-se que os dados podem ser gerados por todos os usuários ativos nas redes sociais; porém, é necessário avaliar a circunstância em que esses dados são gerados. Mesmo em se tratando de eventos reconhecidos e conceituados na área da moda, o volume elevado de dados gerados torna impossível garantir a credibilidade do conteúdo que cada usuário gera. Contudo, no trabalho de descoberta de conhecimento, é possível destacar conteúdos que tiveram repercussões, o que contribui para a avaliação da credibilidade do conjunto de dados. Nessa primeira avaliação, subjetiva, atribuiu-se ao conjunto de dados credibilidade média.
- *Veracidade* - Os dados gerados pelas fontes correspondem à realidade dos fatos ou podem ser oriundos de um evento sem muita importância, ocorrido localmente? O conteúdo dos dados é obtido diretamente da fonte ou existem possibilidades de os mesmos terem sido manipulados? A avaliação dessa característica é semelhante à da credibilidade. O conjunto de dados é referente a um evento importante da indústria da moda, mas a maneira como esses dados foram gerados não favorece o controle para avaliar a veracidade dos comentários, das opiniões e dos sentimentos postados. Nesse caso, tendo em vista a importância e a proporção do evento de moda, o resultado da avaliação da veracidade foi médio, com a ressalva de que a avaliação pode sofrer alterações de acordo com o resultado da descoberta de conhecimento. Os conteúdos (postagens) são oriundos dos usuários da fonte de dados e não foram manipulados.
- *Imparcialidade* - A fonte de dados é isenta de influência? O conjunto de dados em questão não é isento de influência, mas é afetado por ela de maneira diferente da de outros conjuntos de dados em que as consequências podem ser negativas para o produto. Assim, no caso da moda, as influências das mídias sociais, por envolver propagandas e opiniões, são relevantes, pois podem se tornar tendências ou gerar temas para as distintas coleções. A imparcialidade é um ponto importante para o processo de descoberta de conhecimento. Como as influências sobre as fontes podem indicar tendências de moda, a avaliação a respeito da imparcialidade do conjunto de dados foi baixa.
- *Utilidade* - Anteriormente ao levantamento de dados da fonte, é possível verificar sua utilidade para a descoberta de conhecimento? Além das postagens com os comentários dos usuários de redes sociais, apresentam-se outras informações, tais como: marcação de uma postagem como favorito, data de criação da postagem, identificação do usuário, quantidade de compartilhamento, quantidade de *retwetter*, localização do usuário (latitude e longitude), idioma e endereço de URL. Inicialmente, o conjunto de dados apresentou utilidade, mas essa avaliação será confirmada ao longo do processo de descoberta de conhecimento, ao final do qual os dados poderão ser considerados fortemente úteis.
- *Atualizada* - O conteúdo da fonte de dados encontra-se atualizado? O conjunto de dados foi extraído justamente no período em que ocorreram os eventos, portanto, está atualizado.
- *Inconsistência* - O conteúdo da fonte de dados apresenta ruídos? Na primeira análise dos dados, foram constatados muitos ruídos, como erro de digitação, termos e caracteres irrelevantes.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

- *Complexidade* - O conteúdo da fonte de dados apresenta estrutura complexa e necessita de ferramentas intermediárias para transformá-la? As postagens do conjunto de dados avaliado são textuais, portanto, em formato não estruturado. Isso as faz ser consideradas complexas quando comparadas com formatos não estruturados, o que implica que necessitam de técnicas e ferramentas para ser tratadas.
- *Objetividade* - O conjunto de dados corresponde ao objetivo do processo de descoberta de conhecimento? O conjunto de dados contém postagens relacionadas aos eventos de moda analisados e foi extraído exatamente no período em que esses eventos estavam ocorrendo. Portanto, apresenta evidências claras de que corresponde ao objetivo traçado no processo de descoberta de conhecimento.

Etapa 4 – Decisão das Soluções Tecnológicas

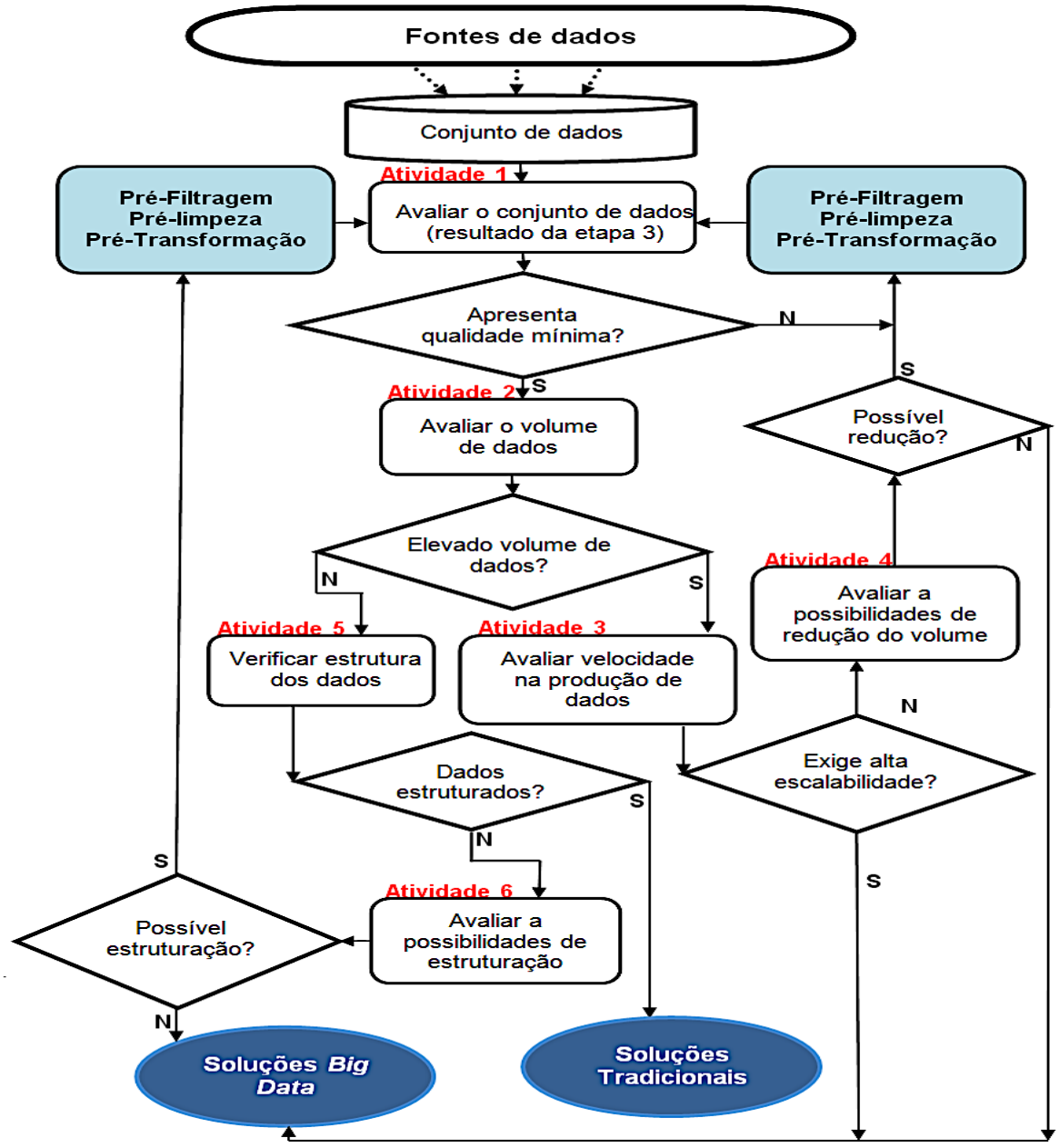
Esta etapa visa o processamento do conjunto de dados bruto e a seleção das soluções tecnológicas. Nesse processamento, busca-se melhorar a qualidade do conjunto de dados em função da solução a ser utilizada. O intuito é melhorar a qualidade e não afastar nenhuma possibilidade de aplicação desse conjunto em soluções tradicionais. Tais soluções estão sendo utilizadas há mais tempo, se comparadas às soluções *Big Data*. Portanto, suas técnicas foram testadas, modificadas e atualizadas, o que proporciona maior confiabilidade em sua utilização.

A avaliação e a manipulação do conjunto de dados para definir a escolha da solução adequada e esgotar todas as possibilidades de aplicar as soluções tradicionais, passaram pelas seguintes atividades representadas na Figura 3:

- Atividade 1 - avaliação do conjunto de dados: a análise prévia realizada no conjunto de dados permitiu avaliar as características apresentadas.
- Atividade 2 - avaliação do volume dos dados: caso a quantidade de dados a serem armazenados seja elevada, apresenta-se a necessidade de redução de seu volume. Notadamente, a fonte de dados *twitter* dispõe de grande variedade e elevado volume de dados; no entanto para essa aplicação, os conjuntos NYFW e MFW, extraídos com o auxílio da linguagem R, não apresentaram volumes elevados.
- Atividade 3 - avaliação da velocidade na produção dos dados: como a extração dos conjuntos NYFW e MFW foi realizada após os eventos, tais conjuntos apresentaram velocidade estática.
- Atividade 4 - avaliação da possibilidade de redução do volume: Os conjuntos de dados NYFW e MFW não apresentaram volume elevado e, por esse motivo, não necessitaram de sistemas com alta escalabilidade.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering*, *7*(3), 106-125.

Figura 3. Diagrama de Atividades



Fonte: Autores.

- Atividade 5 - verificação da estrutura dos dados: como os conjuntos de dados NYFW e MFW foram dispostos em linhas e colunas. Ao considerar o conjunto de dados como um todo, isto é, postagens e atributos, tal conjunto assume formato estruturado. Entretanto, o objetivo é que esta aplicação se restrinja apenas às postagens, em formato de texto escrito e linguagem natural, portanto o conjunto é não estruturado.
- Atividade 6 - avaliação das possibilidades de estruturação: Quando se consideram somente as postagens para realizar a descoberta de conhecimento, indica-se o BD da família NoSQL, que, além de atender aos cenários de análises textuais, possibilita a inclusão de dados advindos de outras fontes e com estruturas diferentes.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

3.1.3 Fase III: Sondagem e Uso de TMT

Esta etapa tem como objetivo preparar os dados e, assim, com foco no armazenamento e\ou mineração, possibilitar a efetiva descoberta de conhecimento. Nessa aplicação, o processo de descoberta de conhecimento realizado nos conjuntos NYFW e MFW evidenciará os termos conforme as frequências com que aparecem nas postagens. Para tanto, foi necessário realizar a preparação desses dados, por meio das seguintes tarefas: *i)* remoção de termos sem relevância por intermédio do pacote (tm) da linguagem R. Além da lista de termos de acordo com o idioma, disponibilizada por esse pacote, foi elaborada outra lista com termos específicos; *ii* padronização de todos os caracteres em caixa baixa; *iii)* remoção de espaços em branco; *iv)* redução dos termos aos seus radicais, como *paraded* para *parade* e *walked* para *walk.*

Decisão das Soluções Tecnológicas

Foi realizada uma avaliação das características dos bancos de dados NoSQL, e assim, foi possível realizar orientações de uso do BD para as fontes de dados úteis no PDP. Portanto, para os conjuntos de dados originados nas redes sociais, os bancos de dados recomendados foram o MongoBD e Cassandra. Para trabalhar com a linguagem R, foram desenvolvidos os pacotes "mongolite" do autor Ooms (2017) e "RCassandra" do autor Urbanek (2015), os quais fornecem interfaces de comunicação e manipulação de dados entre os bancos e a linguagem R.

O processamento com a linguagem R do conjunto de dados NYFW e MFW foi limitado pela RAM (*Random Access Memory*) do computador. Dessa forma, para melhorar o processamento, foi criado um ambiente de armazenamento na linguagem R conectado ao MongoDB por meio do pacote mongolite criado por Ooms (2017). O pacote mongolite que, além de possibilitar a interface com a linguagem R, fornece suporte para indexação, *map-reduce* e funções *streaming*, essencial para conjunto de dados que exige alta escalabilidade. O código de conexão e armazenamento está descrito no Apêndice G.

3.1.4 Fase IV: Descoberta de conhecimento

Na Fase IV, da descoberta de conhecimento, são apresentadas as possibilidades para a realização da análise do conjunto de dados, visando estabelecer a técnica adequada para visualizar tanto os resultados provenientes dessas análises quanto os dados brutos. A execução dessa fase tem a finalidade de produzir resultados que possam produzir conhecimentos novos e úteis. Isso implica uma seleção adequada da fonte de dados e a realização do processo de busca de conhecimento alinhada aos objetivos definidos na Fase I do modelo proposto (diagnóstico).

Análise e técnica de visualização

As técnicas de visualização de informações são utilizadas para simplificar a tarefa de interpretação dos resultados obtidos por meio dos algoritmos empregados na descoberta de conhecimento. Essas técnicas baseiam-se na capacidade humana de percepção para analisar eventos complexos, permitindo reconhecer o que é útil e ao mesmo tempo desconsiderar o que não é de interesse (Rabelo, Dias, Franco e Pacheco, 2008).

O fato de os conjuntos NYFW e MFW apresentarem diferentes atributos dá margem às mais diversificadas análises; no entanto, nesta aplicação, o foco esteve nas postagens de usuários do

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering*, *7*(3), 106-125.

twitter. Foram utilizados algoritmos com cálculos de ocorrência dos termos e, posteriormente, aplicadas ações para associar e/ou agrupar os termos que apresentaram maior relevância.

Após o cálculo de frequência, as técnicas de visualização *Word Cloud* foram apresentadas para a equipe de colaboradores da indústria. Nessa reunião, a análise dos termos destacados permitiu identificar *i)* nome de estilistas, *designers*, atrizes e modelos; *ii)* estação do ano; *iii)* estilo de roupa; *iv)* termos referentes às possíveis tendências; *v)* usuários; *vi)* portais de moda e *vii)* termos aparentemente inúteis para a descoberta de conhecimento. Para preservar as marcas e nomes, as mesmas foram substituídas por termos genéricos.

Figura 4. Visualização *Word Cloud* para os Conjuntos NYFW e MFW

(a) NYFW **(b) MFW**

Fonte: Autores

O cálculo da frequência desses termos foi realizado de acordo com o número de postagens do conjunto de dados. Isto é, a frequência de determinado termo no conjunto de dados correspondeu à divisão da frequência absoluta desse termo pelo total de postagens do conjunto de dados. Tem-se, assim,

$$Fd(x) = \frac{Fa(x)}{Np},$$

em que F_d, F_a e N_p representam a frequência do termo em relação ao número de postagens, a frequência absoluta e o número total de postagens, respectivamente.

Os termos destacados pelos colaboradores da indústria foram submetidos a tarefa de associação por meio da função "*findAssocs*", pertencente ao pacote "tm" da linguagem R e assim gerou o Quadro2 para o conjunto de dados NYFW e o Quadro 3 para o conjunto de dados MFW.

Quadro 2. Frequências da Associação entre os Termos - NYFW

Associação	Termo pesquisado No conjunto de dados NYFW		Termos associados	
	Termos	Frequência	Termo	Frequência
1	*model-x*	6,0%	*designer-x* *designer-y* *atriz-x*	63% 29% 38%
2	*style*	7,2%	*street*	63%
3	*designer-x*	5,2%	*acessdes-y* *model-x*	61% 38%
4	*favorit*	2,0%	*colege* *color*	57% 57%
5	*designer-w*	3,3%	*model-y*	62%
6	*season*	2,3%	*yeezi*	31%

Fonte: Autores

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

Com base no Quadro 2, foram realizadas algumas observações.

- Associação 1 - o termo "*model-x"* refere-se a uma modelo e atriz e está associado aos termos "*designer-x*" e ao "*designer-y*", que se referem a desenhistas norte-americanos. Essa associação indica que as postagens referentes ao desfile da "*model-x*" com roupas do "*designer-x"* tiveram mais destaque, uma vez que apresentam maior frequência. Além disso, evidencia a associação do termo "*model-x*" com termo "*atriz-x*", que representa uma atriz norte-americana de ascendência brasileira.
- Associação 2 - o termo "*style*" está associado ao termo "*street*", indicando que o "estilo de rua" se destacou mais do que outros estilos.
- Associação 3 - o termo "*designer-x*" se destacou no desfile quando associado aos acessórios ("*acessdes-y*") e ao termo "*model-x*", como ocorreu na associação 1.
- Associação 4 - o termo "*favorit*" (favorito) destacou-se quando associado aos termos "*college*" e "*color*", evidenciando o favoritismo das cores colegiais.
- Associação 5 - o termo "*designer-w*" refere-se a um desenhista de moda norte-americano e se destacou quando associado ao termo *"model-y"*, que representa uma modelo e atriz residente no Japão.
- Associação 6 - o termo "*season*" destacou-se quando associado ao termo "*yeezi*", referente a uma nova linha de tênis.

No Quadro 3, mostram-se exemplos de associação entre os termos pesquisados no conjunto MFW e suas frequências em relação aos termos associados

Quadro 3. Frequências da Associação entre os Termos - MFW

Associação	Termo pesquisado no conjunto de dados MFW		Termos associados	
	Termos	**Frequência**	**Termo**	**Frequência**
1	*fw*	26,2%	model-g	35%
2	*style*	6,9%	street	73%
3	winter	6,3%	*theimonation*	55%
4	*collection*	6,7%	*theimonation*	40%
5	*Walk*	6,4%	model-e	35%
6	*new*	5,6%	model-e	48%
7	*Model-e*	2,0%	*Brand-t* *designer-s*	44% 39%
8	*designer-c*	5,5%	*model-b*	36%
9	*trend*	2,4%	*rainbow* *moeeztali*	35% 36%

Fonte: Autores

Com base no Quadro 3, podem ser realizadas algumas observações para cada associação:

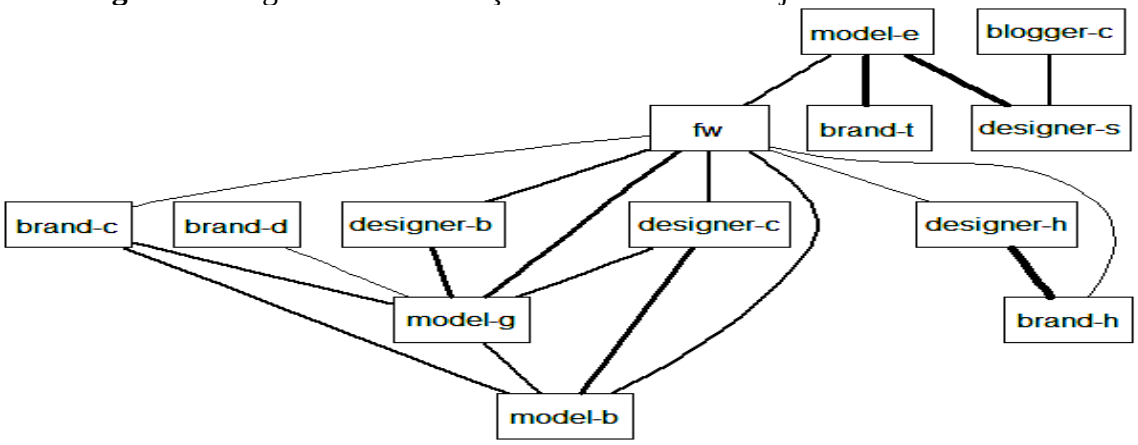
- Associação 1 - o termo "fw" (abreviação de "*fashionweek*" em português "semana da moda") destacou-se quando associado ao termo "*model-g*", indicando que as postagens referentes à modelo norte americana se sobressaíram na semana da moda.
- Associação 2 - o termo "*style*", assim como no evento de moda NYFW, destacou-se quando associado ao termo "*street*", evidenciando o "estilo de rua".
- Associação 3 e 4 - os termos "*winte*r" (inverno) e "*collection*" (coleção) destacaram-se quando associados ao termo "*theimonation*", que se refere a um "*blog*" alternativo, aparentemente desvinculado de marcas.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering*, *7*(3), 106-125.

- Associação 5 e 6 - os termos "*walk*" (andar) e "*new*" (novo) destacaram-se quando associados ao termo "*model-e*", referente a uma modelo sueca;
- Associação 7 - o termo "*model-e*" destacou-se quando associado aos termos "*brand-t*" e "*designer-s*".
- Associação 8 - o termo "*designer-c*", referente a um desenhista italiano, quando associado ao termo "*model-b*", referente a uma modelo americana, indica o destaque das postagens que relatavam que a modelo desfilou com coleções do referido desenhista.
- Associação 9 - o termo "*trend*" (tendência) associou-se aos termos "*rainbow*" (arco iris) e "*moeeztali*". Isso indica o destaque das postagens que relacionavam a tendência a arco íris e o termo "*moeeztali"*, referente a um aplicativo desenvolvido para celulares, utilizado para seguir *blogueiros* relacionados à moda.

A Figura 5 mostra a associação entre os termos referentes a designers, marcas e modelos identificados no conjunto MFW. Nessa visualização, os termos, representados por retângulos, são interligados por linhas de diversas espessuras: quanto mais espessas, maior é o grau de associação entre os termos representado por elas.

Figura 5. Diagrama de Associação de Termos do Conjunto de Dados MFW



Fonte: Autores

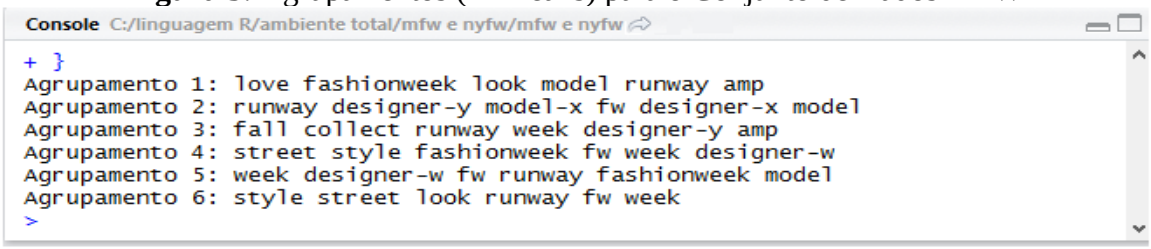
A interpretação da Figura 5 mostra claramente a associação das modelos aos designers e marcas. Por meio da análise realizada na primeira etapa da presente fase, identificou-se que, entre os termos do conjunto MFW, a maior frequência foi para o termo "*model-g*".

De posse dessa informação, foi realizada uma pesquisa sobre a atuação dessa modelo na semana do evento em Milão, constatando-se que ela desfilou para diferentes marcas e *designers*, que, nesta aplicação, são representados pelos termos: "*brand-u*", "*designer-b*", "*brand-x*", "*brand-c*", "*designer-c*" e "*brand-d*". Observa-se que, nas associações ilustradas na Figura 44, os termos "*brand-u*" e "*brand-x*" estão ausentes, o que leva a pressupor que a modelo, representada pelo termo "*model-g*", não se destacou tanto quando foi relacionada às marcas representadas pelos termos "*brand-u*" e "*brand-x*". Todavia, o termo "*model-g*", quando relacionado aos termos "*designer-b*", "*designer-c*" e "*brand-d*", destacou-se com frequência de 29%, 19% e 16%, respectivamente.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering*, *7*(3), 106-125.

Continuando com a análise dos conjuntos NYFW e MFW, especificamente quanto ao procedimento de agrupamento, foram gerados seis grupos (*cluster*), cada qual contendo seis termos, como ilustram as Figuras 5 e 6.

Figura 5. Agrupamentos (K-Means) para o Conjunto de Dados MFW

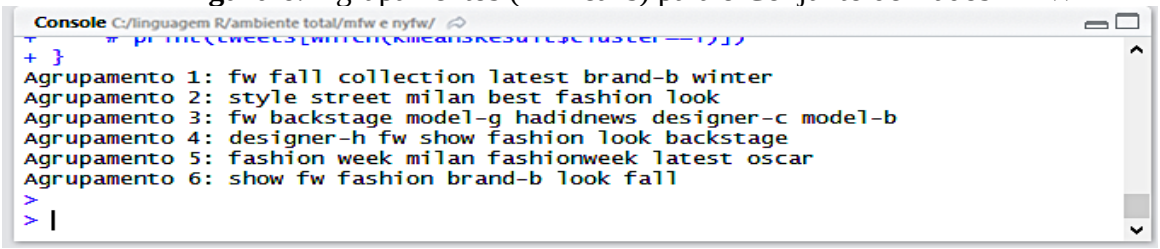


Fonte: Autores

A interpretação da Figura 5, com seus respectivos termos, resultou em algumas observações sobre os agrupamentos (abaixo foram destacados apenas os agrupamento relevantes):

- Agrupamento 2 - os termos destacados nesse agrupamento e confirmados em breve pesquisa levaram à conclusão de que, no desfile realizado em Nova Iorque, a modelo representada pelo termo "*model-x"* utilizou um vestido azul marinho da coleção do desenhista representado pelo termo "*designer-y"*. Porém, em outro dia, no mesmo evento, essa modelo desfilou com um vestido branco, uma capa emparelhada pertencentes à coleção do desenhista representado pelo termo "*designer-x"*.
- Agrupamento 3 - o agrupamento realizado entre os termos indica que a coleção de outono está relacionada ao desenhista representado pelo termo "*designer-y*".
- Agrupamento 4 – indica associação entre a expressão estilo de rua, representada pelo termos "*style*" e "*street*", e o desenhista representado pelo termo "*designer-w*".
- Agrupamento 6 - evidencia que o visual do estilo de rua se sobressaiu nos desfiles ocorridos na semana da moda.

Figura 6. Agrupamentos (K-Means) para o Conjunto de Dados MFW



Fonte: Autores

A interpretação da Figura 6, com seus respectivos termos, resultaram nas seguintes evidencias:

- Agrupamento 1 - identifica semelhança entre os termos "*fw*", "*collection*" (coleção), "*latest*" (recente) e a marca representada pelo termo "*brand-b*", ou seja, indica que a marca citada e sua respectiva coleção se destacaram no evento.
- Agrupamento 2 - evidencia as postagens relacionadas ao estilo de rua. Após rápida pesquisa, buscando pelos termos destacados nesse agrupamento, foi confirmado que portais de moda ressaltaram o estilo de rua;
- Agrupamento 3 - identifica os relacionamentos entre os termos "*model-g*" e "*model-b*", quando associados ao "*designer-c"*. Após rápida pesquisa, descobriu-se que as modelos,

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering*, *7*(3), 106-125.

representadas pelos termos "*model-g*" e "*model-b*", são irmãs e desfilaram vestindo roupas da coleção do desenhista representado pelo termo "*designer-c*".

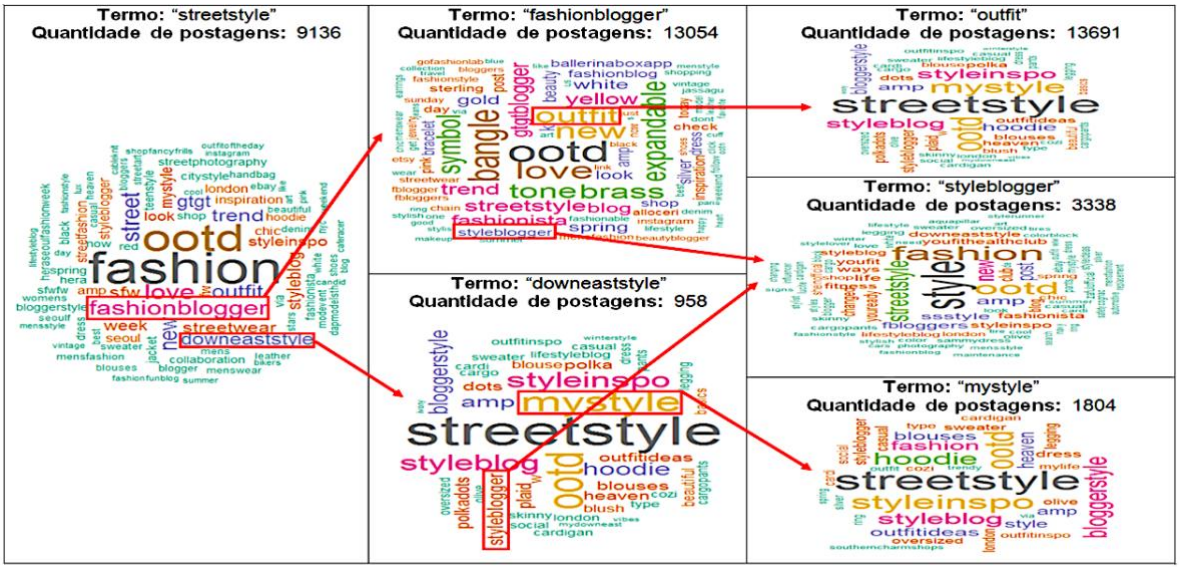
- Agrupamento 4 – destaca a associação entre o termo "*designer-h*" e os termos *"look"* (visual) e *"backstage"* (bastidores), o que denota o visual das roupas desse desenhista nos bastidores do evento.
- Agrupamento 6 - mostra a associação entre os termos referentes ao evento propriamente dito e a marca representada pelo termo "*brand-b*" e o visual de outono representado pelos termos "*look*" e "*fall*".

Os destaques gerados pelos processos de associação e agrupamento apenas demonstraram os padrões de dependência e similaridade entre os termos. Isso permitiu a construção de supostas evidências, mas estas foram analisadas e complementadas com novas pesquisas, de forma a confirmar ou não seu aproveitamento como conhecimento novo e útil.

A utilização de diferentes parâmetros nas tarefas de análise pode produzir novos e diferentes resultados, mas, em razão da limitação do escopo desta aplicação, não foram esgotadas outras possibilidades de associação e agrupamento.

Dessa ideia, aliada à necessidade de se visualizar simultaneamente os termos destacados nesses conjuntos, foi elaborada a Figura 7, que apresenta uma sequência de visualizações *word cloud* interligadas por termos previamente selecionados. Inicialmente, foi gerada a primeira visualização com a junção dos termos "*street"* e "*style"*, disposta mais à esquerda da Figura 7. Na sequência, foram selecionados os termos "*fashionblogger*" e "*downeaststyle*", destacados por retângulos com bordas e setas de cor vermelha, para gerar a visualização dos conjuntos e, assim, sucessivamente, de acordo com a importância do termo a ser escolhido.

Figura 7. Grupo de Visualização *Word Cloud* Ligado por Termos



Fonte: Autores

A dinâmica do processo empregado para gerar a sequência de visualização, mostrada na Figura 7, eleva o volume de dados e, consequentemente, amplia a velocidade na produção, coleta e processamento desses dados. Nesse caso, para a execução da sequência de visualização, torna-se necessária a criação de uma ferramenta com escalabilidade e desempenho satisfatório.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

Embora tenha ocorrido limitação na quantidade dos conjuntos extraídos e em suas respectivas visualizações, algumas evidenciam foram destacadas:

- A visualização do conjunto de dados "*streetstyle*" destacou o termo "*fashionblogger*", que faz referência a "*bloggers"* de moda; nesse subconjunto, destacaram-se os termos "*yellow*" (amarelo) e "*White*" (branco). O termo "*downeaststyle*" refere-se a uma loja sediada nos Estados Unidos que, mantém um *blog* para discutir as tendências da moda e comercializar seus produtos via comércio eletrônico.
- A visualização do subconjunto de dados "*downeaststyle*" destacou os termos "*polka*" (bolinhas), "*plaid*" (xadrex) e "*sweater*" (suéter).
- Na visualização do subconjunto de dados "*outfit*" (roupa) os termos destacados foram, "*polka*" (bolinha), "*hoodie*" (moletom com capuz) e *plaid* (xadrex).
- A visualização do subconjunto de dados "*mystyle*" (meu estilo) evidenciou os termos "*hoodie*" (moletom com capuz) e "s*tyleinspo*", que aparece em outras visualizações e se refere a um portal on-line de moda, cujo objetivo é inspirar o mundo da moda.

Os colaboradores da indústria enfatizaram ainda a importância de se conhecer e acompanhar *blogs* relevantes para o mundo da moda. A rede mundial de computadores tem conseguido juntar em comunidades pessoas geograficamente distantes e com interesses comuns, quebrando tabus relacionados às tendências do mundo da moda.

4. RESULTADOS

O MDC-PDP prevê o armazenamento dos conhecimentos extraídos. No entanto, previamente ao armazenamento, foi necessário confrontá-los com o objetivo estabelecido na fase inicial. Depois de avaliada essa concordância, os prováveis conhecimentos extraídos foram organizados em formato de questionário. Esse questionário apresenta, para cada um dos conjuntos NYFW e MFW, nove prováveis conhecimentos, além de outros cinco relacionados aos termos "*street*" e "*style*".

O questionário foi respondido pelos colaboradores de duas indústrias têxteis. A finalidade foi avaliar a probabilidade de esses conhecimentos se tornarem, de fato, conhecimentos novos e úteis para auxiliar no processo de desenvolvimento da coleção, conforme a definição de Fayyad, Piatetsky-Shapiro e Smyth (1996) de que os conhecimentos extraídos devem ser válidos, novos e potencialmente úteis.

Com base na pesquisa empírica realizada na indústria têxtil, por meio de observação direta e de depoimentos de seus colaboradores, foi possível tecer algumas considerações em relação aos conhecimentos extraídos.

1 As evidências relacionadas aos eventos de moda, auxiliam o direcionamento das pesquisas para o desenvolvimento da coleção.

2 Os colaboradores enfatizaram a importância de se realizar o processo de extração de conhecimento logo após o evento, uma vez que, quanto antes esse conhecimento esteja disponível, maior contribuição será proporcionada.

3 Embora a frequente associação entre os termos "estilo" e "rua", em princípio, não tenha causado surpresa nos colaboradores, sua associação a outros termos destacados pode indicar tendências.

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

4 Diante da elevada quantidade de marcas, modelos e *designers*, os colaboradores destacaram a importância de se acompanhar e avaliar as evidências das postagens publicadas nas redes sociais.

Tais considerações são relevantes e podem até mesmo auxiliar na execução das atividades descritas nas fases e etapas do MDC-PDP, isto é, embasar os parâmetros a serem definidos

No Quadro 4, estão sintetizadas as respostas fornecidas pelos colaboradores da indústria 1.

Quadro 4. Respostas ao Questionário – Indústria 1

Conjunto de dados	Quantidade de conhecimentos	Novo	Útil	Novo e útil
NYFW	9	5	7	5
MFW	9	3	7	3
street e *style*	5	3	5	3

Fonte: Autores

Para que se tornem efetivamente válidos para auxiliar o desenvolvimento da coleção, os conhecimentos extraídos devem ser simultaneamente novos e úteis para todos os colaboradores diretamente envolvidos na criação da coleção. Em relação a esse requisito, o Quadro 4 mostra que, dentre as respostas fornecidas pelos colaboradores da indústria 1 para os questionamentos referentes ao conjunto NYFW, cinco atestaram positivamente.

Dentre as nove respostas referentes aos questionamentos do conjunto MFW, apenas três foram positivas. O mesmo número de respostas positivas foi dado para os questionamentos referentes aos termos *"street*" e "*style*".

Além dos conhecimentos extraídos terem sido avaliados pelos colaboradores da indústria 1, com intuito de ampliar a relevância desses conhecimentos, foi realizada também uma avaliação por colaboradores de outra indústria têxtil. Essas avaliações estão sintetizadas no Quadro 5.

Quadro 5. Respostas ao Questionário – Indústria 2

Conjunto de dados	Quantidade de conhecimentos	Novo	Útil	Novo e útil
NYFW	9	5	5	3
MFW	9	4	8	4
street e *style*	5	2	5	2

Fonte: Autores

Evidentemente, a relevância dos conhecimentos extraídos é inerente às avaliações realizadas individualmente pelas indústrias estudadas. No entanto, quando se tratou de avaliar se concomitantemente os conhecimentos eram novos e úteis, tanto por parte dos colaboradores da indústria 1 quanto da indústria 2, observou-se que houve consenso na avaliação de alguns conhecimentos, mais especificamente 34% dos conhecimentos extraídos.

Em virtude da natureza exploratória das análises realizadas na Fase IV, naturalmente surgem indagações em relação aos questionamentos atestados como novos e úteis. Nesse caso, faz-se necessário um estudo mais detalhado para entender a origem e os motivos que levaram essas postagens a se destacar no *twitter*.

No processo de descoberta de conhecimento, a importância está no direcionamento que o conhecimento extraído pode fornecer para ampliar o campo de visão e gerar novas concepções

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

para o desenvolvimento da coleção. Dessa maneira, os conhecimentos extraídos podem não ser a solução para o objetivo traçado, mas podem ser evidências que levam à sua consecução.

5. CONCLUSÃO

Sabe-se que os dados são fator determinante e fundamental nos processos de negócio de uma organização, da mesma forma que sua qualidade. Em face disso, o MDC-PDP vai ao encontro das necessidades das organizações manufatureiras, cujo interesse é, por meio dos recursos de dados, reforçar as decisões que serão utilizadas no PDP. Com essa finalidade, na Fase II do modelo proposto, voltada aos dados e suas respectivas fontes, empregou-se a análise de faceta e se adotaram procedimentos para verificar, trabalhar, acompanhar e garantir essa qualidade.

Com a aplicação do modelo proposto, foi possível evidenciar que esforços empreendidos na compreensão antecipada dos dados podem ocasionar redução da complexidade dos dados extraídos e tornar o *Big Data* viável para uso na indústria.

Ao longo do desenvolvimento da aplicação do MDC-PDP, chegou-se a algumas constatações: *i)* o uso do modelo proposto mostrou-se adequado para tornar, os colaboradores da indústria, próximos do processo de descoberta de conhecimento. Essa constatação ocorre principalmente na Fase II, cuja finalidade é avaliar e compreender o conjunto de dados e sua respectiva fonte; *ii)* a utilização do modelo proposto garante a construção de um roteiro específico para cada objetivo traçado no processo de descoberta de conhecimento e, assim, favorece a criação de ambientes tecnológicos, por meio da utilização e/ou desenvolvimento de ferramentas.

Embora as soluções tradicionais e *Big Data* disponham de ferramentas para realizar a análise e dar suporte ao tratamento de dados, o fator humano é imprescindível para sua manipulação. Outro entendimento, não menos importante, é o da possibilidade de dissociação entre volume e valor dos dados, isto é, o entendimento de que o valor dos dados não está vinculado ao seu volume.

Uma das recomendações para a continuidade desta pesquisa é a análise das incertezas e do interesse do setor industrial pela implantação do processo de descoberta de conhecimento. Isso vale também para as indústrias de pequeno porte, uma vez que, existem uma diversidade de ferramentas tecnológicas gratuitas para a utilização e/ou desenvolvimento de outras ferramentas que contribuam, do início ao fim, para a aplicação do MDC-PDP. Dessa forma, o maior investimento será concentrado nos profissionais.

6. REFERÊNCIA

Asamoah, D. A., & Sharda, R. (2015). Adapting CRISP-DM Process for Social Network Analytics: Application to Healthcare. *Twenty-First Americas Conference on Information Systems*, Puerto Rico, 2015, (Mdd), 1–12.

Begoli, E., & Horey, J. (2012). Design Principles for Effective Knowledge Discovery from Big Data. *Proceedings of the Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture*, Helsinki, Finland, 2012. 215–218. https://doi.org/10.1109/WICSA-ECSA.212.32

Davenport, T. (2014). *Big data at work: dispelling the myths, uncovering the opportunities*. Harvard Business Review Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Citação (APA): Rabelo, E., Campos, F. C. de, & Silva, L. M. C. da. (2021). Aplicação de um modelo de descoberta de conhecimento na era do *Big Data*. *Brazilian Journal of Production Engineering, 7*(3), 106-125.

Gantz, J., Reinsel, D., & Shadows, B. D. (2012). *The Digital Universe in 2020*. IDC iView "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," 2007(December 2012), 1–16.

Halvorsen, K., Hoffmann, J., Coste-Manière, I., & Stankeviciute, R. (2013). Can fashion blogs function as a marketing tool to influence consumer behavior? Evidence from Norway. *Journal of Global Fashion Marketing, 4*(3), 211–224. https://doi.org/10.1080/20932685.2013.790707

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115. https://doi.org/10.1016/j.is.2014.07.006

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In System Sciences (HICSS), *Proceedings of the Hawaii International Conference*, HI, USA, 46.

Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big Data in product lifecycle management. *The International Journal of Advanced Manufacturing Technology, 81*(1–4), 667–684. https://doi.org/10.1007/s00170-015-7151-x

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.pdf

McAfee, A., & Brynjolfsson, E. (2012). *Big Data. The management revolution*. Harvard Buiness Review, 90(10), 61–68. https://doi.org/10.1007/s12599-013-0249-5

Milonas, E. (2011). *Wittgenstein and web facets*. NASKO, 3(1), 33–40.

Piatetsky, B. G. (2014). CRISP-DM, *still the top methodology for analytics, data mining, or data science projects*. Retrieved from http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Rabelo, E., Dias, M., Franco, C., & Pacheco, R. C. S. (2008). Information visualization: Which the most appropriate technique to represent data mining results? *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation, CIMCA 2008*, Vienna, Áustria. (pp. 1228–1233). https://doi.org/10.1109/CIMCA.2008.139

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, 181*, 526-534.

Shiri, A. (2014). Making sense of big data: A facet analysis approach. *Knowledge Organization, 41*(5), 357–368. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84913553153&partnerID=tZOtx3y1

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Zhang, Y., Ren, S., Liu, Y., & Si, S. (2017). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *Journal of Cleaner Production, 142*, 626–641. https://doi.org/10.1016/j.jclepro.2016.07.123

Zhuang, Y., Wang, Y., Shao, J., Chen, L., Lu, W., Sun, J., … Wu, J. (2016). D-Ocean: an unstructured data management system for data ocean environment. *Frontiers of Computer Science, 10*(2), 353–369. https://doi.org/10.1007/s11704-015-5045-6

Zikopoulos, P., Eaton, C., & et al. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.