Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 09/11/2019; 1st round notif.: 05/01/2020; New version: 18/02/2020; 2nd round notif.: 23/03/2020
Camera ready: 11/05/2020; Edition review: 04/06/2020; Available online: 02/07/2020; Published: 02/07/2020

Tendências de Pesquisas em Mineração de Dados Educacionais em MOOCs: um Mapeamento Sistemático

Research Trends in Educational Data Mining in MOOCS: A Systematic Mapping of Literature

Vanessa Faria de Souza
Programa de Pós-Graduação em Informática na Educação - Universidade Federal do Rio Grande do Sul (UFRGS)
vanessa.souza@ibiruba.ifrs.edu.br

Gabriela Trindade Perry
Programa de Pós-Graduação em Informática na Educação - Universidade Federal do Rio Grande do Sul (UFRGS)
gabriela.perry@ufrgs.br

Resumo

Massive Open Online Courses (MOOCs) são uma modalidade de curso que utilizam plataformas online e atraem diferentes perfis de estudantes, ofertando oportunidades de qualificação - seja formal ou informal - em um formato bastante dinâmico. Uma característica das plataformas que ofertam tais cursos é a capacidade de armazenar uma grande quantidade de dados, o que possibilitou a exploração destes dados por meio de técnicas de Mineração de Dados Educacionais (MDE). Nesse contexto, foi conduzido um mapeamento sistemático de literatura em cinco bases de dados com o propósito de verificar quais as vertentes de estudos em destaque quanto ao uso de MDE em MOOCs. A busca compreendeu o período de 2015 a 2020, sendo que 158 foram selecionados. Os resultados revelaram que estudos relativos à Análise de Comportamento, Predição (de Desempenho, de Abandono, de Conclusão), Mineração de Texto e Sistemas de Recomendação são os mais frequentes. Além disso, foram identificadas áreas com potencial de exploração, como Social Network Analysis (SNA), Digital Learning Ecosystem (DLE) e Análise de Mind Wandering (MW). Ademais, foram levantados métodos e ferramentas usados nas pesquisas, bem como desafios do uso de MDE na pesquisa sobre MOOC. Conclui-se que a questão do desequilíbrio de classes, provocado pela baixa adesão aos cursos, é um dos maiores desafios.
Palavras-Chave: *MOOCs, Mineração de Dados Educacionais, Mapeamento Sistemático.*

Abstract

Massive Open Online Courses (MOOCs) use online platforms and attract different student profiles, offering qualification opportunities - whether formal or informal - in a very dynamic format. A characteristic of the platforms that offer such courses is the ability to store a large amount of data, which made it possible to explore it through Educational Data Mining (EDM) techniques. In this context, a systematic scoping was conducted in five databases with the purpose of discover research trends regarding to the use of EDM in MOOCs. The search covered the period from 2015 to 2020, selecting 158 papers. The results revealed that studies related to Behavior Analysis, Prediction (Performance, Abandonment, Conclusion), Text Mining and Recommendation Systems are the most frequent. Promissing researches were also identified, such as Social Network Analysis (SNA), Digital Learning Ecosystem (DLE) and Mind Wandering Analysis (MW). Methods and tools used in research were listed, as well as challenges in the use of EDM in research on MOOC. We concluded that the issue of class imbalance, caused by low adherence to courses, is one of the biggest challenges.
Keywords: *MOOCs, Educational Data Mining, Systematic Scoping.*

De Souza, V. F., & Perry, G. T. (2020). Research Trends in Educational Data Mining in MOOCS: A Systematic Mapping of Literature (Tendências de Pesquisas em Mineração de Dados Educacionais em MOOCs: um Mapeamento Sistemático). Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 28, 491-508. DOI: 10.5753/RBIE.2020.28.0.491

1 Introdução

Durante a década de 90, houve uma importante evolução em relação às tecnologias empregadas no âmbito da Educação a Distância, representada pela popularização dos Ambientes Virtuais de Aprendizagem (AVAs) – dentre os quais o Moodle, Teleduc e o Rooda. No final de 2011 surgem os Massive Open Oline Courses (MOOCs), os quais compreendem um tipo de Curso aberto ofertado por meio da utilização de AVAs, que apresentam um novo cenário para a EAD, no que se refere à transição da lógica da transmissão para a lógica da comunicação entre os mais diversos perfis de usuários (Wagner et al, 2016). A maior parte da oferta de MOOCs está concentrada no cenário educacional americano, tendo como principais expoentes Coursera, Udacity e EDX (Wagner et al., 2016).

Para Butcher (2014) os MOOCs possuem duas características principais que os distinguem em comparação aos cursos on-line tradicionais: (1) acesso aberto[1]: de forma que qualquer pessoa, mesmo não estando matriculada em uma instituição de ensino participe; e (2) escalabilidade: devem suportar uma grande quantidade de participantes. Neste sentido, em função da variedade de recursos de aprendizagem, do elevado número de inscritos e da capacidade das plataformas armazenarem dados de navegação, abrem-se novas possibilidades de compreensão do aprendizado a partir de uma perspectiva quantitativa (Lee, 2018). Por consequência, a análise das ações dos estudantes é uma tarefa fundamental para detectar barreiras para a aprendizagem, principalmente em MOOCs - que costumam apresentar baixas taxas de conclusão (He et al., 2015).

Nesse cenário, a Mineração de Dados Educacionais (MDE), uma área de pesquisa interdisciplinar que lida com o desenvolvimento de métodos para explorar dados originados no âmbito educacional (Romero & Ventura, 2016) tem ganhado destaque. As técnicas de MDE almejam a extração de informações dos dados registrados pelas plataformas no decorrer da realização de um MOOC, e que podem conduzir à identificação de características comportamentais e indicadores relacionados à aprendizagem (Lu et al., 2017). Baker, Isotani & Carvalho (2011) afirmam que as principais contribuições da MDE são: (1) a criação de modelos para melhor compreender os processos de aprendizagem; e (2) o desenvolvimento de métodos mais eficazes para dar suporte à aprendizagem quando o aluno estuda utilizando softwares educacionais ou Ambiente Virtuais de Ensino Aprendizagem (AVAs).

Em vista disso, esta pesquisa teve o objetivo de realizar um mapeamento sistemático da literatura, a fim de levantar estudos com enfoque na aplicação de MDE em MOOCs, de modo a identificar: (1) as tendências temáticas e propósitos gerais de sua utilização em cursos dessa natureza; (2) as técnicas e ferramentas mais utilizadas; (3) as oportunidades de pesquisa e (4) os principais desafios da aplicação dessas técnicas.

2 Mineração de Dados Educacionais

Quando é preciso analisar uma grande quantidade de dados, é imprescindível contar com recursos computacionais, caso contrário a tarefa torna-se impraticável. Sendo assim, necessita-se de ferramentas que auxiliem na tarefa de verificar, interpretar e relacionar esses dados, com o objetivo de gerar conhecimento útil e relevante – o que, segundo Los Reyes et al. (2019) já era um objetivo das técnicas de Mineração de dados (MD), empregadas para identificar padrões de comportamento e encontrar insights que gerem melhorias em produtos e serviços. No que tange

[1] Acesso *Aberto* não significa especificamente gratuito.

ao uso destas técnicas em problemas e contextos relacionados à educação, os objetivos são semelhantes. No caso específico da aplicação em problemas relacionados à Educação, Romero e Ventura (2007) elencam questões que diferenciam a aplicação de Mineração de Dados Educacionais da mineração em outros domínios:

1) Objetivos: que podem se relacionar à pesquisa (a) aplicada, que busca responder questões práticas, por exemplo: como melhorar o processo de aprendizagem e (b) pura, que objetiva por exemplo dar sentido às observações. Na maioria das vezes esses objetivos são difíceis de quantificar e exigem seu próprio conjunto especial de técnicas de medição.

2) Dados: em ambientes educacionais, existem muitos tipos diferentes de dados disponíveis para mineração. Esses dados são específicos da área educacional e, portanto, possuem informações semânticas intrínsecas, relacionamentos com outros dados, e vários níveis de hierarquia significativa.

3) Técnicas: problemas educacionais têm algumas características especiais que exigem que a questão da mineração seja tratada de uma maneira diferente. Embora, a maioria das técnicas tradicionais de mineração possam ser aplicadas diretamente, outras não podem e devem ser adaptadas ao problema educacional específico. Um exemplo é que em se tratando de cenários comuns, de mineração de dados, as variáveis são em sua maioria numéricas, tratáveis diretamente por algoritmos de *Machine Learning*, enquanto que em ambientes educacionais a grande maioria é categórica, o que implica esforço em pré-processamento para codificar essas variáveis em numéricas, para que então possam ser interpretadas pelos algoritmos.

Todos estes objetivos são importantes em qualquer cenário educacional, porém, em um curso online, com muitos alunos, e que (na maioria) não têm acompanhamento de professores ou tutores, técnicas de análise de dados em massa tornam-se a solução mais viável. Em suma, a MDE abrange desenvolvimento, pesquisa e aplicação de métodos para detectar padrões em grandes conjuntos de dados educacionais, que de outra forma seria difícil ou impossível analisar devido ao seu enorme volume (Romero & Ventura, 2010; 2013).

Por meio do uso da MDE em MOOCs, talvez seja possível acompanhar e compreender o processo de aprendizagem, bem como outros fatores que influenciam a aprendizagem. Por exemplo, talvez seja possível identificar que tipo de abordagem instrucional (e.g. aprendizagem individual ou colaborativa) proporciona mais benefícios ao aluno, observando variáveis que representem seu engajamento com o curso. Além disso, abre-se a possibilidade de verificar se o aluno está aprendendo ou confuso, identificar níveis de motivação, envolvimento nas atividades on-line, descoberta de elementos ou indicadores comportamentais de conclusão e sucesso em um curso, identificar padrões de interação, descobrir estratégias que contribuam para a permanência dos estudantes (Pursel et al., 2016, apud Martin & Piovesan, 2019), elementos que podem ajudar a personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem (Baker, Isotani & Carvalho, 2011).

No decorrer dessa pesquisa, artigos que apresentam revisões de literatura sobre MDE foram encontrados: Sukhija, Jindal & Aggarwal (2015); Shahiria, Husaina & Rashida (2015) e Aldowaha, Al-Samarraiea & Fauzy (2019). A revisão de Sukhija, Jindal & Aggarwal (2015) foca na evolução do MDE, abrangendo o período de 2001 a 2015, tendo sido identificadas cinco lacunas na área: (1) indisponibilidade de conjuntos de dados consistentes que sejam grandes o suficiente para refletir o sistema educacional e seu funcionamento; (2) integração e versatilidade nos conjuntos de dados; (3) grande parte da técnicas de mineração foram aplicadas isoladamente e poucos trabalhos foram realizados utilizando técnicas híbridas; (4) falta de confiança nos resultados da MDE; (5) necessidade de comparar métodos. A revisão sistemática de Shahiria, Husaina & Rashida (2015) teve como objetivo fornecer uma visão geral das técnicas de mineração de dados que são usadas para prever o desempenho dos alunos, apontando

que a maioria dos pesquisadores utilizou a média acumulada de notas e a avaliação interna nos conjuntos de dados, enquanto para técnicas de previsão, os métodos de classificação Rede Neural e Árvore de Decisão foram os mais usados.

Finalmente, a pesquisa realizada por Aldowaha, Al-Samarraiea e Fauzy (2019), revisou 402 artigos publicados de 2000 até 2017, abordando quatro temas a respeito de MDE e Learning Analytics: Aprendizagem Suportada por Computador, Análise Preditiva Suportada por Computador, Análise Comportamental Suportada por Computador e Análise De Visualização Suportada por Computador. Os autores relatam que a aplicação de técnicas como clustering, regras de associação, mineração visual de dados e testes estatísticos diversos podem proporcionar benefícios significativos e, portanto, instar as instituições de ensino superior a adotá-los onde factível. Além disso, também argumentam que a aplicação de MDE e Learning Analytics no ensino superior pode ajudar a desenvolver uma oferta mais focada no aluno e fornecer dados e ferramentas que as instituições serão capazes de usar para previsão em tempo real. Tais pesquisas forneceram informações substanciais sobre a base teórica, metodológica e objetivos desse campo em rápido crescimento. Tais revisões não têm foco em MOOCs.

Ademais, foram encontradas revisões que enfocam MOOC: Fournier, & Kop (2015); Davis et al. (2018); Duru, Dogan & Diri, (2016) – porém estas não tratam sobre MDE (de forma direta). Fournier, & Kop (2015) realizaram uma revisão sistemática sobre as várias estruturas que visam orientar os esforços de pesquisa, desenvolvimento e avaliação em torno MOOCs, revelando temas de pesquisa e desenvolvimento, tais como: (1) análise de aprendizado; (2) big data; (3) mineração de dados educacionais e (4) questões de ética e privacidade em ambientes de rede e (5) o uso de dados pessoais de aprendizado para alimentar o processo de pesquisa e desenvolvimento.

O trabalho realizado por Davis et al. (2018), oferece uma síntese de pesquisas publicadas entre 2009 e 2017, que realizaram avaliações empíricas de estratégias de aprendizagem, e por meio de uma busca sistemática encontraram 126 artigos e os categorizaram de acordo com as estratégias de aprendizado apresentadas. Como resultados os autores identificaram três estratégias mais promissoras para alavancar efetivamente o aprendizado em MOOCs: (1) Aprendizado Cooperativo, (2) Simulações/Jogos e (3) Multimídia Interativa.

Enfim, Duru, Dogan & Diri, (2016) apresentam uma revisão de literatura com foco em pesquisas que realizaram análise de desempenho e aprendizado em MOOCs. Os autores enumeram descobertas sobre o uso da previsão de desempenho de alunos e Learning Analytics em MOOCs, salientando que as áreas mais pesquisadas nesse segmento são: (1) Previsão de resultados acadêmicos, (2) Painéis de cursos e programas, (3) Avaliação Curricular, (4) Priorização de resultados de aprendizado, (5) Definição de políticas de curso e instrução e (6) Definição de qualidade acadêmica.

Percebeu-se que tanto mapeamentos como revisões sistemáticas de literatura com foco na aplicação de Mineração de dados Educacionais em MOOCs não são numerosas, e investigações nesse campo podem representar uma contribuição para pesquisadores que tenham interesse nesta área de estudo.

3 Procedimentos Metodológicos

Para realização deste mapeamento, adotou-se um método usado em revisões sistemáticas de literatura, pois minimiza o enviesamento da escolha da literatura, na medida em que é feita uma busca dos textos publicados sobre o tema em questão (Denyer & Tranfield, 2009). Ramos, Faria & Faria (2014) indicam que o propósito de uma revisão sistemática é resumir a melhor pesquisa

disponível acerca de uma questão específica, o que é feito por meio da síntese dos resultados de diversos estudos. Já um mapeamento, como o próprio nome indica, não busca avaliar como o conjunto da literatura responde determinadas questões de pesquisa, e sim fazer um apanhado geral sobre determinada área, apresentando um panorama que permita identificar oportunidades de pesquisa. Como em uma revisão sistemática, além do mais deve-se utilizar procedimentos bem definidos para encontrar, avaliar e sintetizar os resultados de pesquisas relevantes na área em estudo – porém o faz com mais abrangência.

Para a condução deste mapeamento utilizou-se como referência Ramos, Faria & Faria (2014) que propõem o seguinte protocolo: (1) definir os objetivos; (2) definir as strings de busca; (3) definir bases de dados; (4) definir critérios de inclusão; (5) definir critérios exclusão; (6) definir critérios de validade metodológica; (7) tabular os dados e (8) realizar tratamento de dados. Ressalta-se que Ramos, Faria & Faria (2014) apresentam este protocolo como método para conduzir revisões sistemática, mas consideramos que ele poderia ser adaptado para um mapeamento. De acordo com Ramos, Faria & Faria (2014), é imprescindível que sejam registradas todas as etapas de pesquisa, não só para que esta possa ser replicável por outro investigador, como também para se aferir que o processo em curso segue uma série de etapas previamente definidas e respeitadas. Os autores propõem, neste domínio, que se implementem os passos do protocolo ilustrado na Figura 1.

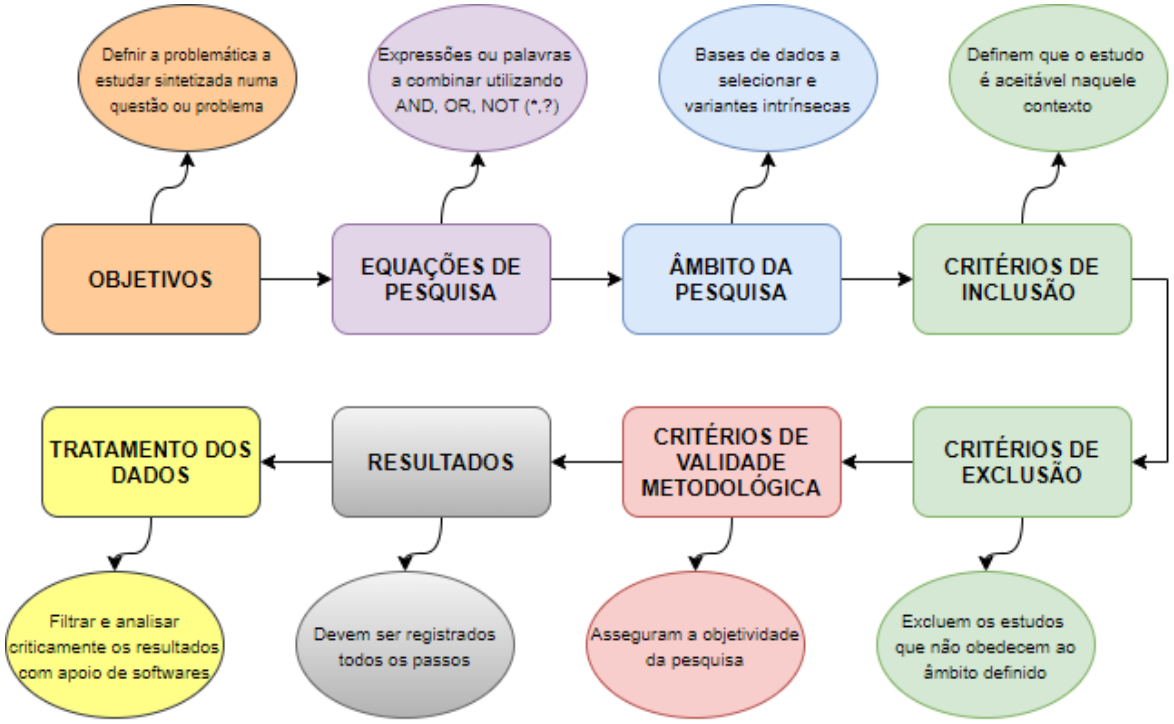

Figura 1: Etapas do Processo de Revisão Sistemática de Literatura.
Fonte: Adaptado de Ramos, Faria & Faria (2014).

O primeiro passo do processo é a definição dos Objetivos da Pesquisa, que neste trabalho é responder à 4 questões de pesquisa:

QP1. Quais as tendências temáticas?

QP2. Quais as técnicas e algoritmos mais utilizados?

QP3. Quais as principais oportunidades de pesquisa?

QP4. Quais os desafios mais relatados?

Com relação a String de pesquisa, optou-se por utilizar um termo genérico: ((“educational data mining” OR “EDM”) AND (“massive open online course” OR “MOOC”)), e seu equivalente em português, a fim de alcançar um amplo conjunto de estudos.

O âmbito da pesquisa corresponde as bases de dados nas quais os artigos foram buscados, no presente estudo: IEEE Explore Digital Library, ACM Digital Library, ERIC, Science Direct, e Taylor e Francis. Elas foram selecionadas por indexarem muitos trabalhos internacionais, que têm maior alcance. Além do que, os dois principais eventos e periódicos de MDE são indexados nessas bases a *“International Conference on Educational Data Mining”* e o *“Journal of Educational Data Mining”*.

Quanto aos critérios de inclusão e exclusão definiu-se: (1) Inclusão – (a) em Inglês ou Português; (b) completos; (c) que tenham passado por um processo de revisão cega; (d) e publicados entre 2015 e 2020. (2) Exclusão – (a) artigos duplicados; (b) revisões ou mapeamentos de literatura; (c) resumos ou artigos publicados em congressos nacionais; (d) publicados em revistas que não sejam *peer review*. Decidiu-se não incluir artigos apresentados em congressos nacionais ou regionais pois entender-se que nestes fóruns as pesquisas tendem a estar em um estágio inicial, sem resultados completos. Além disso, exclui-se, por óbvio, artigos escritos em idiomas não compreendidos pelos autores ou que não atendessem à temática “MDE em MOOCs”.

O sexto passo, relativo aos critérios de validade metodológica, devem assegurar a replicação do processo. Para este trabalho todos os artigos retornados da busca passaram por três níveis de triagem: (1) Os resumos de todos os artigos retornados, na busca, foram lidos pela primeira autora, dos quais foram selecionados aqueles que atendem aos critérios de inclusão mencionados, e alguns de forma aleatória foram lidos pela segunda autora, para validar se os critérios de inclusão e exclusão estavam sendo seguidos da maneira correta; (2) Depois, todos os artigos que passaram pela primeira triagem, tiveram seus resumos analisados pela primeira autora, com a finalidade de gerar as categorias temáticas; (3) Por fim, com o intuito de verificar se as pesquisas selecionadas realmente satisfaziam aos critérios de inclusão, ou deveriam ser descartados por atender a algum critério de exclusão, a primeira autora realizou a leitura dos artigos, enfocando a metodologia utilizada pelos autores, resultados alcançados e suas conclusões. A leitura da revisão bibliográfica e introdução foi feita de forma rápida, apenas se não fosse compreendido algo nos demais itens uma leitura mais apurada era conduzida.

Nesta etapa também, como descrito, foi efetuada a geração de categorias temáticas de pesquisa, sendo empregado o procedimento “Keywording” (geração de palavras chave). Keywording consiste na tarefa de leitura dos resumos dos artigos com o propósito de encontrar palavras-chave que contribuam para gerar as categorias temáticas, e caso o resumo não forneça informações suficientes o procedimento alternativo é recorrer à introdução e conclusão Petersen et al. (2008, apud Souza et al, 2019).

No tocante à Tabulação e ao Tratamento dos dados, o mapeamento foi conduzido com o apoio das ferramentas Parsifal[2] e Excel. Os dados foram analisados em uma dimensão quantitativa (contagem de ocorrência para elaboração de estatísticas descritivas) e qualitativa (categorização). Todos os resumos e elementos essenciais dos artigos, incluídos nesse mapeamento, estão disponíveis no link que consta no rodapé[3]. O detalhamento da condução do mapeamento, baseado no protocolo proposto por Ramos, Faria & Faria (2014), pode ser observado na Figura 2, na qual são apresentados todos os passos metodológicos realizados,

[2] https://parsif.al/
[3] https://docs.google.com/spreadsheets/d/1FPu4Cr39zWHU0JsmHapeAjXfFXOHDv0X/edit#gid=1746542574

desde a definição dos objetivos até o tratamento dos dados, com o propósito de obter resultados baseados nas quatro questões de pesquisa formuladas. Na seção 4, os resultados são apresentados e as questões são respondidas.

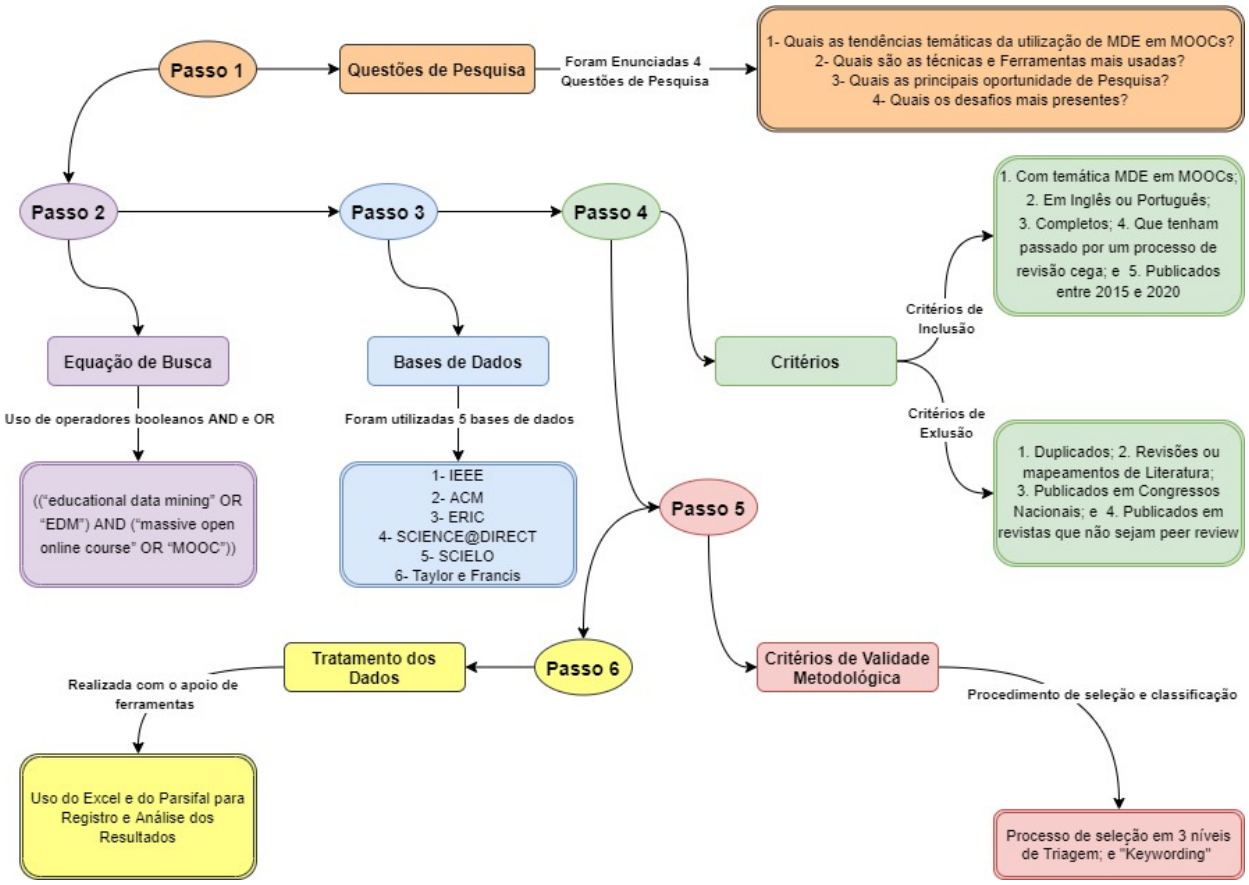

Figura 2: Detalhamento da Condução da Revisão.

4 Resultados

O mapeamento sistemático abrangeu o período de 2015 a 2020, sendo selecionados artigos completos que relataram o uso de MDE em MOOCs. A Tabela 1 exibe a quantidade de trabalhos retornados dos Passos 2 até 5. No primeiro nível da triagem (resultado da string) foram selecionados 376 artigos, que em seguida foram avaliados quanto aos seus critérios de inclusão e exclusão, resultando em 158 trabalhos.

Tabela 1: Totais de pesquisas encontradas e selecionadas.

Base	Quantidade de Artigos Retornados	Quantidade de Artigos Selecionados
ACM Digital Library	96	28
IEEE Digital Library	104	67
Science@Direct	69	14
SciElo	24	0
ERIC	56	44
Taylor e Francis	27	5
TOTAL	**376**	**158**

Salienta-se que as pesquisadoras, após realizarem a busca nas bases de dados, perceberam que muitos artigos cujas temáticas também eram classificadas (nas palavras-chave) como *"Learning Analytics" e "Machine Learning"* envolvem Mineração de dados Educacionais. Por este motivo, buscou-se em Siemens & Baker (2010), Liñán & Pérez, (2015) e Moissa, Gasparini & Kemczinski (2015) esclarecimentos sobre diferenças e similaridades entre estes dois termos, para definir se poderiam ser diferenciados de forma objetiva.

Liñán & Pérez, (2015) usam o artigo de Siemens & Baker (2010) como base, para apresentar as diferenças entre as áreas. Segundo os autores, as diferenças são em relação à automatização, escopo, origens, métodos e objetivos. Em relação à automatização, argumentamos que em nenhuma destas áreas pode-se prescindir do conhecimento a respeito do problema, como pode-se pensar que as nomenclaturas "aprendizagem de máquina" e "mineração de dados" denotam. Qualquer algoritmo de "aprendizagem" de máquina demanda, pelo menos, a seleção de variáveis – na maior parte das vezes demanda também a preparação dos dados, a modelagem e a própria montagem/configuração do algoritmo.

Desta forma, não se pode usar "automatização" como critério de diferenciação. Em relação ao escopo ("sistemas", para Learning Analytics e "componentes" para EDM), argumenta-se que esta definição demanda uma referência – por exemplo, um plugin do Moodle é um componente ou um sistema? Ademais, acredita-se que não se pode fazer uma diferenciação em termos de objetivos, como, por exemplo: prever comportamentos, encontrar padrões e comparar amostras, pois estes dizem respeito às perguntas de pesquisa, e não ao método.

Finalmente, as origens de cada uma das áreas não são um critério útil para diferenciar uma pesquisa. Moissa, Gasparini & Kemczinski (2015) fizeram um mapeamento sistemático para responder à esta dúvida, e argumentam que estas áreas podem ser diferenciadas em termos de processo e atores: enquanto Learning Analytics é um ciclo que análise que se retroalimenta, o processo de EDM é sequencial e eventualmente chega ao seu final.

No entanto, entendemos que isso implica fazer uma diferenciação a posteriori, ou seja, apenas seria possível identificar uma área ou outra após a pesquisa ser concluída (para avaliar se ela foi "sequencial e chegou ao fim" ou se "retroalimentou"). As próprias autoras, ao concluir sobre o resultado, ponderam que verificaram que as duas áreas possuem definições e objetivos similares. Portanto, como não foi possível encontrar critérios que as diferenciassem, estas áreas foram tratadas como similares.

A fonte de publicação com a maior quantidade de artigos retornados e selecionados foi a *International Conference on Educational Data Mining*, com 40 artigos. Em seguida, as revistas *IEEE Access*, *Computers in Human Behavior e Computers & Education*, respectivamente com 5, 5 e 4 publicações. A quantidade de artigo por base foi: IEEE, 67; Eric, 44; ACM, 28; Science Direct, 14; Taylor & Francis, 5. Os países de origem mais frequentes (considerando apenas primeiros autores) foram Estados Unidos (47) e China (40). Espanha, Suíça, Japão, Austrália, Índia, Equador, Brasil, Alemanha, Chile e Hungria tinham entre 7 e 2 artigos. As publicações sobre o tema tiveram seu ápice entre 2016 e 2018 – a série que vai de 2015 a 2020 indica 20, 38, 38, 40, 19 publicações.

A síntese dos resultados do mapeamento pode ser vista no Quadro 1, que agrupa os artigos em função das tendências temáticas, listando: quantidade de artigos, principais algoritmos e técnicas, oportunidades de pesquisa e desafios mais relatados. Este quadro lista apenas as

temáticas com pelo menos 10 artigos[4]. As temáticas mais presentes somaram 110 artigos, 69% do total.

Quadro 1: Síntese dos Resultados Alcançados, agrupados pela temática.

TEMÁTICA (QTD DE ARTIGOS)	OPORTUNIDADES DE PESQUISA	DESAFIOS
Predição: de Desempenho (18) de Abandono (16) de Conclusão (6)	Implementação de modelos de alunos que levem em consideração a leitura como um componente fundamental da aprendizagem Aplicação do método de empilhamento – abordagem para combinar várias técnicas de aprendizado de máquina em um modelo preditivo – é conhecido por ter um bom desempenho com dados desequilibrados	Desequilíbrio de classe e grande variabilidade no formato de dados MOOCS. Como a maioria dos alunos não conclui ou avança no curso, os algoritmos supervisionados aprendem de forma tendenciosa. Demanda muito esforço na fase do Pré-processamento.
Análise de Comportamento (30)	Conclusão de curso, procrastinação, regularidade e interação. Exploração de variáveis relacionadas à exibição de vídeos, autoteste e interação no fórum. Análise de cliques em relação ao tempo na tarefa, para avaliar engajamento	“Cold start problem” ou "arranque a frio" é um problema potencial em sistemas de modelagem automatizada de dados, diz respeito à questão de que o sistema não pode extrair inferências para usuários ou itens sobre os quais ainda não reuniu informações suficientes. O Fluxo de Cliques pode ser uma medida de análise ruim, pois alunos podem somente clicar na tarefa, mas não estarem engajados em sua realização
Mineração de Texto (28)	Análise da polaridade de palavras de um texto, com o objetivo de entender seu efeito sobre processos e estados de aprendizagem afetiva. Integração de fontes de informação externa ao curso, isso permite uma análise mais detalhada do comportamento do aluno em atividades acadêmicas que devem ser cumpridas	Na Mineração de texto, diferentes gramáticas e diferentes estilos linguísticos tornam a aplicação de algoritmos uma tarefa complexa, e os principais softwares de mineração de texto são proprietários e por isso têm pouca flexibilidade, e estão disponíveis na maioria das vezes para o inglês.
Sistemas de Recomendação (12)	Uso de algoritmos de agrupamento e redes neurais.	Desequilíbrio de classe e grande variabilidade no formato de dados MOOCS.

Além destas 4 temáticas, também encontrou-se pesquisas sobre: Análise de Redes Sociais (6), Identificação de Trapaças (6), Aprendizagem Auto Regulada (5), Análise de Sentimentos (4), Análise de Vídeos (4), Desenvolvimento de Plataformas MOOC (3), Avaliação por Pares (2), Gerenciamento de Personalização (2), Análise de Pré-Requisito (2), Sistemas de Gerenciamento de Dados (2), Análise de Trajetórias de Aprendizagem (2), Análise de Currículo (1), Análise de Design (1), Digital Learning Ecosystem (1), Análise do Engajamento (1), Geração de Usuários Artificiais (1), Análise de Feedback (1), Gamificação (1), Análise da Motivação (1), Mind Wandering (1), Análise de Dados de Recursos Educacionais (1).

[4] Corte arbitrário, definido pelas autoras.

4.1 QP1 - Quais as tendências temáticas e propósitos gerais da utilização de MDE em MOOCs?

Foram identificadas 25 temáticas em meio aos artigos pesquisados, sendo que 4 delas (apresentadas no quadro 1) respondem por 69% das pesquisas, sendo a Predição (de Desempenho, Abandono e Conclusão) o tema mais investigado. A previsão de Desempenho procura identificar com antecedência como será a performance do aluno no decorrer do curso, para poder intervir caso necessário e assim melhorar seu processo de aprendizagem. Por exemplo, Waheeda et al. (2020) implementam uma rede neural artificial profunda (DNN), em um conjunto de dados extraídos do fluxo de cliques de uma plataforma MOOC, para prever o desempenho de estudantes e assim poder auxiliar aqueles que estejam em risco. Sobre a Predição de Abandono estudos nessa linha objetivam conhecer alunos que pretendem desistir antes do encerramento do curso, como no trabalho de Wen et al (2020), os autores utilizam uma rede neural convolucional (CNN), que supera o desempenho de outros métodos mais tradicionais na previsão do abandono e em posse dessas previsões tutores, professores e gestores podem realizar ações que diminuam os índices de desistência. Sobre Predição de Conclusão, Pigeau, Aubert & Prié, (2019) apresentam um estudo de caso sobre um conjunto de dados fornecido pelo OpenClassrooms, que são modelados de 8 formas diferentes, usando algoritmos de classificação e abordagens baseadas em sequência, como mineração de padrões de processo.

Análise de Comportamento é a segunda categoria mais numerosa em quantidade de publicações. Entre os trabalhos que selecionamos para apresentar como modelo deste tipo de pesquisa, está o de Lan et al. (2017) que propõem um modelo de aprendizado que relaciona o comportamento ao assistir vídeos e o envolvimento com o desempenho em atividades avaliativas. A maioria dos trabalhos com essa tendência tem foco no melhoramento da experiência educacional dos MOOCs.

A terceira categoria mais numerosa é a Mineração de Texto, que engloba Análise em Fóruns de Discussão, um tema muito recorrente. A Análise de Fóruns de Discussão possui vários propósitos, dentre os quais - detecção de erros dos alunos, relevância temática, engajamento, postagens que necessitam da atenção dos professores. Um exemplo é Guo et al. (2019) que apresentaram uma nova rede neural híbrida para identificar postagens “urgentes” que requerem atenção imediata dos instrutores em fóruns de discussão. Geralmente, essas duas vertentes costumam possuir características interligadas, pois muitos estudos em fóruns foram elaborados por meio da mineração de texto. Além disso, foram encontrados artigos que tratam da mineração de texto em e-mail, redes sociais – como em Joksimović et al. (2015), no qual realizaram análises de tópicos de discurso em Redes Sociais para descobrir o que alunos de MOOCs postam.

O tópico Sistemas de Recomendação também é frequente e diz respeito à implementação de sistemas que fazem alguns tipos de sugestões para usuários de MOOCs, como por exemplo: recomendar contatos que possuam características semelhantes, recomendar conteúdos e também cursos dentro das plataformas. Em Labarthe et al. (2016) os autores tentam aprimorar a experiência do MOOC com um sistema de recomendação que fornece a cada aluno uma lista individual de contatos com alto potencial de congruência.

4.2 QP2 - Quais as técnicas e algoritmos mais utilizados?

Muitas técnicas diferentes são empregadas em pesquisas sobre MDE, e elas baseiam-se principalmente na aplicação de algoritmos que podem detectar informações relevantes em meio a muitos dados. Nas pesquisas sobre Análise de Comportamento e Sistemas de Recomendação, foram citados algoritmos de aprendizagem de máquina bastante conhecidos como KNN, Clustering, Support Vector Machines, Naive Bayes, Random Forest, Decision Tree e Apriori,

bem como modelos como Alocação latente de Dirichlet e Regressão logística/linear. Quando a temática era Predição (Desempenho, Abandono e Conclusão), Sistemas de Recomendação e Mineração de Textos, notou-se maior uso de Redes Neurais dos mais diversos tipos: artificiais, profundas, recorrentes e convolucionais. Notou-se uma predominância de algoritmos altamente especializados em pesquisas sobre Mineração de Texto.

Como exemplo, cita-se Balint (2016), que analisou o comportamento de resolução de problemas de física de alunos matriculados em um MOOC, por meio do algoritmo de agrupamento K-means. Além do mais, Al-Shabandar et al. (2017) que teve como objetivo a previsão do desempenho em MOOCs, e utilizaram vários algoritmos: Regressão Logística, Latent Dirichlet Allocation (LDA), Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Redes Neurais (tipo - MLP) e Mapa Auto-Organizado (SOM). Cita-se também Kashyap & Nayak (2018), que utilizaram os algoritmos Naïve Bayes, Random Forest, Decision Tree e Support Vector Machine (SVM), para prever o abandono em MOOCs. A Tabela 2 lista os 10 algoritmos mais empregados, considerando as temáticas abrangidas e sua funcionalidade.

Tabela 2: Algoritmos de mineração de dados mais utilizados.

Algoritmo	**Quantidade de Temáticas Abrangidas**	**Funcionalidade**
Clustering (K-means e Hierárquico)	13	Agrupar objetos semelhantes.
Regressão Logística	8	Produz, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável binária.
Support Vector Machine	7	Classifica determinado conjunto de pontos de dados que são mapeados para um espaço de características multidimensional usando uma função kernel
Decision Tree	5	Fornece ao usuário final uma fácil interpretação e desenha uma espécie de caminho a ser percorrido para alcançar um determinado objetivo
K-Nearest Neighbours	5	Determina o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento
Redes Neurais	5	Realiza o aprendizado de máquina bem como o reconhecimento de padrões, atingindo uma solução generalizada para uma classe de problemas
Latent Dirichlet Allocation (LDA)	5	Descreve um conjunto de observações como uma mistura de categorias distintas. É mais comumente utilizado para descobrir um número de tópicos especificado pelo usuário
Naïve Bayes	4	Utiliza dados históricos para prever a classificação de um novo dado.
Random Forest	4	Cria muitas árvores de decisão, de maneira aleatória, formando o que pode-se enxergar como uma floresta, onde cada árvore será utilizada na escolha do resultado final

Em relação às linguagens de programação mais usadas (em ordem alfabética), Python e R foram as mais citadas. As ferramentas de software mais citadas foram (em ordem alfabética) Matlab, NodeJS, RapidMiner, SAS, SPSS, TensorFlow e Weka. Além disso, foram citadas API para redes sociais, como WikiPedia Miner, Kwitty e Facebook3. O Mechanical Turk, serviço da Amazon, também foi utilizado. Em relação à mineração de textos, foi encontrada uma variedade muito significativa de bibliotecas de código, em função da grande especialização das tarefas.

Foram encontrados trabalhos que usam Python e R, como Sunar et al. (2018) e Cobos & Olmos (2018). Sunar et al. (2018) usaram estas linguagens para tipificar os diferentes padrões de comportamento social dos participantes durante um curso, e testaram estatisticamente se existe uma correlação entre a conclusão do curso e os comportamentos modelados, com a finalidade de

entender melhor o engajamento social dos alunos em uma plataforma MOOC e o impacto do engajamento na conclusão do curso. Cobos & Olmos (2018) implementaram um complexo modelo de predição chamado EDX-MAS que possui dois módulos: (1) Módulo de Importação – permite extrair, limpar, selecionar e pré-processar os dados do curso para detecção de abandono e para selecionar atividades relevantes na coleta de dados, além disso, suporta a criação de variáveis de entrada e gerenciamento de armazenamento de dados, codificado na linguagem Python; (2) Módulo de geração de modelo – este módulo suporta a geração de modelos preditivos do curso selecionados por dia ou por semana, codificado em R, utiliza 10 algoritmos de aprendizagem de máquina - Boosted Logistic Regression, Random Forest, Stochastic Gradient Boosting, Naïve Bayes, Extreme Gradient, Boosting Neuronal Network (one hidden layer), Support Vector Machines, Bayesian Generalized Linear Model, K-Nearest Neighbours, Classification e Regression Tree

Ademais, duas ferramentas que já trazem os algoritmos implementados e são amplamente utilizadas em mineração de dados são constatadas, Weka e RapidMiner. Weka é um software livre para mineração de dados, desenvolvido em Java, que se consolidou como instrumento de mineração de dados mais utilizado por estudantes e professores de universidades, muito por causa da facilidade em aplicá-lo. Seu objetivo é agregar algoritmos provenientes de diferentes abordagens/paradigmas de inteligência artificial dedicados especialmente a aprendizagem de máquina. A pesquisa realizada por Brooks, Thompson & Teasley (2015) é um exemplo de sua aplicação. Nela os autores descrevem uma modelagem de alunos com base nos dados coletados dos AVAs, onde implementam modelos preditivos com Decision Tree (tipo J48) usando a ferramenta Weka, com a intenção de prever o desempenho dos alunos. Em relação ao RapidMiner, é um sistema comercial também para análise de dados que utiliza aprendizagem de máquina, muito semelhante ao Weka. No trabalho desenvolvido por An, Krauss & Merceron (2017) os autores o utilizaram na aplicação de algoritmos de Clustering com o intuito de analisar comportamentos típicos de alunos em MOOCs.

4.3 QP3 - Quais as principais oportunidades de pesquisa identificadas?

Dentre as oportunidades listadas no Quadro 1, avalia-se que algumas são mais proeminentes, pois ainda existem poucos trabalhos publicados e há muito espaço para evolução. Por exemplo, no âmbito da Análise de Redes Sociais, Brinton et al. (2018) usaram o conceito de Aprendizagem Social como plano de fundo para analisar as conexões realizadas nos MOOCs e assim avaliar a eficiência dos fóruns de discussão. Brinton et al. (2018) asseguram que a proliferação dos MOOCs apresentou uma infinidade de possibilidades para pesquisas em torno Aprendizagem em redes Sociais.

De mesmo modo, destaca-se a aplicação dos conceitos de Ecossistema de Aprendizagem Digital, que consiste em espécies, populações e comunidades interagindo entre si e com o ambiente. Segundo Galileo et al. (2019) esse modelo pode ajudar educadores e designers instrucionais a reunir informações baseadas em dados dos alunos, esses dados permitem projetar estratégias e atividades inovadoras usando ferramentas diversas, que maximizam a aprendizagem em ambiente *e-learning*.

Por último, destaca-se a Análise de *Mind Wandering* (MW), expressão que traduzimos livremente como “devaneio”. Este conceito abrange a ideia de períodos em que a atenção e o conteúdo dos pensamentos se afastam da ideia original ou da atividade que está sendo executada no momento em que esses pensamentos ocorrem. Essa temática pode ser abordada de diferentes formas, como em Hutt et al. (2017), que investigam o uso de rastreamento ocular no nível de usuário (alunos) para detectar automaticamente este devaneio, enquanto assistem a uma videoaula do MOOC. Os resultados apresentados pelos autores mostraram que esta detecção é

exequível no contexto de assistir a uma videoaula, sendo possível alcançar precisão de 47%. Como pode ser observado, neste Mapeamento, ainda há poucas iniciativas como essa no contexto de MOOCs, dessa forma trabalhos nesse sentido podem ter muito a contribuir com a área.

4.4 QP4 - Quais os desafios mais relatados nesses trabalhos?

Em referência aos principais desafios de pesquisas na área, pode-se apontar como um dos mais recorrentes o Desequilíbrio de Classe, relatado como principal desafio em artigos de 4 temáticas diferentes, tendo sido citado em pelo menos 20 publicações. Uma das causas desse desbalanço é a baixa percentagem de atividades concluídas – desta forma, em abordagens supervisionadas os algoritmos acabam aprendendo de forma tendenciosa, pois existem mais dados rotulados como desistentes do que como concluintes. Em Xing et al. (2016) esse desafio é citado e eles implementam uma solução para conseguir bons resultados, aplicando o método de empilhamento (de algortimos e modelos).

Outro desafio é a diversidade dos dados coletados em MOOCs, que é verificada em especial na predição de conclusão, mas aparece em várias publicações de outras categorias. É um problema que torna difícil a utilização de métodos estatísticos ou de agrupamento simples para criar um modelo preditivo. Kórösi et al. (2018) apresentam uma abordagem de MDE para analisar os dados de fluxo de cliques dos alunos para prever a conclusão em MOOCs, e mencionam essa como a maior dificuldade encontrada para realização do processo, sendo necessária a aplicação de mais de 10 algoritmos para chegar na precisão desejada.

Finalmente, apresenta-se como um obstáculo para os pesquisadores a confiabilidade do fluxo de cliques dos alunos, pois muitos estudantes clicam em uma tarefa, mas não estão engajados em sua resolução. Deve-se, portanto, relacionar o clique do aluno com a quantidade de tempo que este permaneceu em cada tarefa, para validar seu engajamento, e desse modo conseguir informações mais confiáveis, o que requer uma etapa extra de pré-processamento dos dados. He et al. (2018), por exemplo, desenvolveram uma investigação em dados de alunos de MOOCs para extrair padrões de ritmos de aprendizagem, e perceberam que ao utilizar apenas os dados de cliques dos alunos não estavam conseguindo os resultados esperados e com confiabilidade.

5 Considerações Finais

Um grande marco no processo evolucionário da educação foi o surgimento dos MOOCs, que atendem às demandas de um novo cenário tecnológico global. O surgimento desse gênero de curso contribuiu para fortalecer as mudanças nos paradigmas educacionais existentes, além de vir ao encontro do processo de sustentabilidade da educação e dos anseios de um novo perfil de aluno da era digital, cada vez mais presente nas instituições de ensino. Esses cursos foram considerados por Wulf et al. (2014) o próximo passo em educação à distância no mundo. Além dessas características, um dos maiores diferenciais dos MOOCs é a quantidade e a diversidade de dados gerados pelos alunos nas plataformas de oferta, fato que tem oportunizado a exploração dessa massa de dados, para a descoberta de conhecimentos novos a respeito de como os indivíduos estudam, aprendem e interagem.

Nesse contexto, realizou-se um mapeamento sistemático da literatura, com foco em Mineração de Dados Educacionais em MOOCs, pois acredita-se que representa um conjunto de técnicas adequadas para análise de grandes volumes de dados educacionais.

Convém resgatar, a revisão sistemática desenvolvida por Sukhija, Jindal & Aggarwal (2015) que teve como foco a evolução do MDE, abrangendo o período de 2001 a 2015. Nesta pesquisa foram identificadas cinco lacunas: (1) indisponibilidade de conjuntos de dados consistentes que sejam grandes o suficiente para refletir o sistema educacional e seu funcionamento; (2) necessidade de integração e versatilidade nos conjuntos de dados; (3) grande parte da técnicas de mineração foram aplicadas isoladamente e poucos trabalhos foram realizados utilizando técnicas híbridas; (4) há falta de confiança das autoridades nos resultados da MDE e (5) necessidade de comparar métodos. Pode-se dizer, que, embora as descobertas dos autores fossem fortemente fundamentadas, o cenário se modificou bastante deste então. No que se refere à primeira lacuna ressalta-se que com a evolução dos MOOCs, bases com milhões de dados estão disponíveis, como exemplo, pode-se aludir ao trabalho desenvolvido por Northcutt, Ho & Chuang (2016) que utilizaram uma base (plataforma MOOC) com 1.893.092 de usuários que geraram em média de 200 a 1500 interações com a plataforma por curso realizado, portanto, a indisponibilidade de conjuntos de dados já não se configura mais como um problema. No que tange à segunda lacuna, pode-se dizer que em relação à integração das bases, ela se mantém, pois não é possível integrar duas bases de forma simples, sem necessidade de um grande esforço de pré-processamento. Em relação à versatilidade dos dados – qualidade de não ser colinear, ou seja, dos dados não estarem relacionados – pode-se dizer que houve mudança, pois, vários tipos diferentes de dados são usados nos modelos. A terceira lacuna, sobre uso de métodos mistos, pode ser considerada a que mais não coincide com a realidade dos experimentos realizados na área de MDE atualmente. A título de exemplo pode-se mencionar a pesquisa de Xing et al. (2016), onde os autores aplicam o método de *stacking*, no qual utilizam mais de um algoritmo juntos, na identificação precoce de estudantes em risco de desistência. Segundo o levantamento feito no presente artigo, constatamos que cerca de 80% dos trabalhos usam mais de uma técnica de mineração. Com relação à quarta e à quinta lacuna, o presente levantamento não encontrou indícios de mudança no cenário.

Retoma-se também, a revisão de Fournier & Kop (2015), que tinha enfoque exclusivo em MOOCs. Nesta revisão, sobre as várias estruturas que visam orientar os esforços de pesquisa, desenvolvimento e avaliação em torno desses cursos, o documento revela áreas atuais e futuras de pesquisa e desenvolvimento, incluindo: (1) Análise de Aprendizado; (2) Big Data; (3) Mineração de Dados Educacionais; (4) Ética e Privacidade em ambientes de rede e o uso de dados pessoais de aprendizado para alimentar o processo de pesquisa e desenvolvimento. Essas tendências puderam ser confirmadas no decorrer dessa revisão, pois a (1) Análise de Aprendizado permeia as principais temáticas enumeradas neste estudo como Análise de Comportamento, Predição de Desempenho, Predição de Abandono, Predição de Conclusão e Mineração de Texto, bem como as demais encontradas, sendo uma área crucial para o desenvolvimento e avanço dos MOOCs. Além do mais, se pode constatar que os métodos de Mineração de Dados se encaminham para consolidação em investigações conduzidas em MOOCs; essas estão presentes em todos os artigos analisados no decorrer desse processo de revisão, evidenciando sua importância nesse cenário. No que se refere à Ética e Privacidade, uma parcela de 22% dos pesquisadores citou que encaminhou solicitações de acesso aos dados dos estudantes, e os demais, abrangendo a totalidade dos artigos selecionados, salientaram que a identidade dos participantes permaneceu anônima; percebe-se dessa forma que a privacidade é um elemento que tem recebido muita atenção nesses estudos, contudo, pouco se falou sobre questões éticas relacionadas.

Em suma, a revisão implementada apontou que a Análise do de Comportamento, Predição (de Desempenho, de Abandono e de Conclusão), a Mineração de Texto e Sistemas de Recomendação são as temáticas mais abordadas em se tratando da aplicação de Mineração de Dados Educacionais em MOOCs. Destacam-se também as técnicas de mineração mais empregadas nos estudos avaliados, que correspondem aos seguintes algoritmos de classificação e regressão: Clustering, Regressão Logística e Support Vector Machine. Para aplicação dessas técnicas são necessárias, em muitas situações, a utilização de ferramentas computacionais que auxiliam no processo, dessa forma foi possível também mapear as principais ferramentas utilizadas – Pyton, R, Weka e RapidMiner. Além dos temas predominantes, constatou-se que novas abordagens e propósitos de uso de MDE em MOOCs estão emergindo, os quais podem ser consideradas como grandes oportunidades de pesquisa, como no campo da *Social Network Analysis* (SNA), o *Digital Learning Ecosystem* (DLE), e a Análise de *Mind Wandering* (MW). Evidenciou-se também alguns dos principais desafios enfrentados na área, como o desequilíbrio de classes, a complexidade e diversidade (heterogeneidade) dos dados coletados em MOOCs e os dados do fluxo de cliques que muitas vezes não são totalmente confiáveis.

O uso do de MDE em MOOCs evoluiu muito de 2015 a 2020, desta forma salienta-se a importância dos dados educacionais on-line para o avanço na área da educação, da mesma forma que métodos de extração de informações relevantes desses dados, como as técnicas de MDE, que oportunizam conhecer os diferentes perfis e estilos de aprendizagem dos estudantes, e identificar como os conteúdos/recursos educacionais e as ações comportamentais impactam na aprendizagem.

Referências

Aldowaha, H.; Al-Samarraiea, H. & Fauzyb, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics, 37, 13-49.* doi: https://doi.org/10.1016/j.tele.2019.01.007. [GS Search]

Al-Shabandar, R.; Hussain, A.; Andy Laws, A.; Keight, R.; Lunn, J. & Radi, N. (2017). Machine learning approaches to predict learning outcomes in Massive open online courses. *Proceedings of the International Joint Conference on Neural Networks (IJCNN).* Anchorage, AK, USA. doi: 10.1109/IJCNN.2017.7965922. [GS Search]

An, T.; Krauss, C. & Merceron, A. (2017). Can Typical Behaviors Identified in MOOCs Be Discovered in Other Courses? *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*. Wuhan, China, p.220-225. [GS Search]

Baker, R. S. J.; Isotani, S. & Carvalho, A. M. J. B. (2011). Mineração de dados educacionais: oportunidades para o Brasil. *Revista Brasileira de Informática na Educação.* v. 19, n. 2, p. 1-12. doi: http://dx.doi.org/10.5753/rbie.2011.19.02.03. [GS Search]

Balint, T. A. (2016). Depth Analysis of Problem-Solving Profiles of Students in Open Online Environments. *ProQuest LLC, Ph.D. Dissertation, The George Washington University.* p. 1-174. [GS Search]

Brinton, C. G.; Buccapatnam, S.; Zheng, L.; Cao, D.; Lan, A. S. & Felix M. F. (2018) On the Efficiency of Online Social Learning Networks. *Journal IEEE/ACM Transactions on Networking.* v. 26, n. 5, p. 2076-2089. doi: 10.1109/TNET.2018.2859325. [GS Search]

Brooks, C.; Thompson, C. & Stephanie Teasley. (2015). A time series interaction analysis method for building predictive models of learners using log data. *Proceedings of the Fifth*

International Conference on Learning Analytics And Knowledge. p. 126–135. doi: https://doi.org/10.1145/2723576.2723581. [GS Search]

Butcher, N. (2014). Technologies in Higher Education: mapping the terrain. [online]. New York: Unesco. Disponível em: https://iite.unesco.org/files/anons/19/Foresigh_in_ICT_in_HE_BackgroundDocument.pdf. [Acessado 08 Nov. 2019].

Cobos, R. & Olmos, L. (2018). A Learning Analytics Tool for Predictive Modeling of Dropout and Certificate Acquisition on MOOCs for Professional Learning. *Proceedings of the International Conference on Industrial Engineering and Engineering Management (IEEM)*. Bangkok, Thailand, p. 1533-1537. doi: 10.1109/IEEM.2018.8607541. [GS Search]

Davis, D.; Chen, G.; Hauff, C. & Houben, G. J. (2018). Activating learning at scale: A review of innovations in online learning strategies. *Computers & Education, 125, 327-344*. doi: https://doi.org/10.1016/j.compedu.2018.05.019. [GS Search]

Denyer, D. & Tranfield, D. (2009). Producing a systematic review. In: BUCHANAN, D. A.; BRYMAN, A. (Ed.). *The SAGE handbook of organizational research methods*. London, SAGE, p. 671-689. [GS Search].

Duru, I.; Dogan, G.; Diri, B. (2016). An overview of studies about students' performance analysis and learning analytics in MOOCs. *Proceedings of the International Conference on Big Data (Big Data)*. Washington, DC, USA. doi: 10.1109/BigData.2016.7840786. [GS Serch]

Fournier, H. & Kop, R. (2015). MOOC Learning Experience Design: Issues and Challenges. *International Journal on E-Learning, 14(3), 289-304*. doi: http://www.editlib.org/p/150661/. [GS Search]

Galileo, M. M; Roca, M.; Barchino, R.; Hernández, R.; Amado-Salvatierra, H. R. (2019). Applying a Digital Learning Ecosystem to Increase the Effectiveness of a Massive Open Online Course. *Proceedings of the IEEE Learning With MOOCS (LWMOOCS)*. Milwaukee, WI, USA. doi: 10.1109/LWMOOCS47620.2019.8939636. [GS Search]

Guo, S. X.; Sun, X.; Wang, S. X.; Gao, Y. & Feng J. (2019). Attention-Based Character-Word Hybrid Neural Networks With Semantic and Structural Information for Identifying of Urgent Posts in MOOC Discussion Forums. *Journal IEEE Access*. v. 7, p. 120522-120532. doi: 10.1109/ACCESS.2019.2929211. [GS Search]

He, J.; Bailey, J.; Rubinstein, B. I. P. & Zhang, R. Identifying. (2015). At-Risk Students in Massive Open Online Courses. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. p. 1749–1755. [GS Search]

He, J.; Men, C.; Fang, S.; Du, Z.; Liu, J. & Li, M. (2018). Analysis of MOOC Learning Rhythms. *Proceedings of the 20th International Conference on High Performance Computing and Communications*. Exeter, United Kingdom, p. 1555-1562. d*oi:* 10.1109/HPCC/SmartCity/DSS.2018.00255. [GS Serach]

Hutt, S.; Hardey, J.; Bixler, R.; Stewart, A.; Risko, E. & D'Mello, S. K. (2017). Gaze-Based Detection of Mind Wandering during Lecture Viewing. *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*. Wuhan, China, p. 25-28. [GS Search]

Joksimović, S.; Kovanovic, V.; Jovanović, J.; Zouaq, A.; Gasevic, D. & Hatala, M. (2015). What do cMOOC participants talk about in social media?: a topic analysis of discourse in a

Cmooc. *Proceedings of the LAK '15 International Conference on Learning Analytics And Knowledge*. P. 156-165. doi: https://doi.org/10.1145/2723576.2723609. [GS Search]

Kashyap, A. & Nayak, A. Different Machine Learning Models to Predict Dropouts in MOOCs. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Bangalore, India, p.80-85. doi: 10.1109/ICACCI.2018.8554547. [GS Search]

Labarthe, H.; Bouchet, F.; Bachelet, R. &Yacef, K. (2016). Does a Peer Recommender Foster Students' Engagement in MOOCs? *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*. Raleigh, NC, p. 418-423. [GS Search]

Lan, A. S.; Brinton, C. G.; Yang, T. Y & Chiang, M. Behavior-Based Latent Variable Model for Learner Engagement. *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*, Wuhan, China, p. 64-71. [GS Search]

Lee, Y. (2018) Using Self-Organizing Map and Clustering to Investigate Problem Solving Patterns in the Massive Open Online Course: An Exploratory Study. *Journal of Educational Computing*, v. 57, no 2, p. 471-490. doi: https://doi.org/10.1177/0735633117753364. [GS Search]

Liñán, L. C. & Pérez, A. A. J. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time Evolution. *Journal of Educational Technology in Higher Education*. v. 12, n. 3, p. 98-112. doi: 10.7238/rusc.v12i3.2515. [GS Search]

Lu, X.; Wang, S.; Huang, J.; Chen, W.; & Yan, Z. (2017). What Decides the Dropout in MOOCs? *In: DATABASE Systems for Advanced Applications. Cham: Springer International Publishing*. p. 316–327. doi: 10.1007/978-3-319-55705-2_25. [GS Search]

Martin, P. C. & Piovesan, S. D. (2019). Análise da acessibilidade nos MOOCs das universidades federais do Brasil em conformidade com os requisitos do W3C e eMAG. Renote: revista novas tecnologias na educação. v. 17, 3, p. 51-60. doi: https://doi.org/10.22456/1679-1916.99425

Moissa, B.; Gasparini, I. & Kemczinski, A. (2015). Educational Data Mining versus Learning Analytics: estamos reinventando a roda? Um mapeamento sistemático. *Proceedings of the XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015)*, Brasil, p. 1167-1176. doi: http://dx.doi.org/10.5753/cbie.sbie.2015.1167. [GS Search]

Northcutt, C. G.; Ho, A. D. & Chuang, I. L. (2016). Detecting and preventing "multiple-account" cheating in massive open online courses. Jounal Computers & Education. v. 100, p. 71-80. doi: https://doi.org/10.1016/j.compedu.2016.04.008. [GS Serach]

Petersen, K.; Feldt, R.; Mujtaba, S. & Mattsson, M. (2008). Systematic Mapping Studies in Software Engineering. *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, BCS Learning & Development Ltd., Italy, p. 68–77. [GS Search]

Pigeau, A.; Aubert, O. & Prié, Y. (2019). Success Prediction in MOOCs: A Case Study. *Proceedings of the 12th International Conference on Educational Data Mining (EDM)*. Montreal, Canadá, p. 390-395. [GS Search]

Ramos, A.; Faria, P. M.; Faria, A. (2014). Revisão sistemática de literatura: contributo para a inovação na investigação em Ciências da Educação. Revista Diálogo Educ., Curitiba, v. 14, n. 41, p. 17-36. doi: http://dx.doi.org/10.7213/dialogo.educ.14.041.DS01. [GS Search]

Romero, C. Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* v. 1, n. 33, p. 135–146. doi: https://doi.org/10.1016/j.eswa.2006.04.005 . [GS Search]

Romero, C. Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *Jounal IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. v. 40, n. 6, 601-618. doi: 10.1109/TSMCC.2010.2053532. [GS Search]

Romero, C. & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. v. 3, n. 1, p. 12-27. doi: 10.1002/widm.1075. [GS Search]

Romero, C. & Ventura, S. (2016). Educational data science in massive open online courses. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. v. 7, n. 1, p. 1-20. Doi: https://doi.org/10.1002/widm.1187. [GS Search]

Shahiria, A. M; Husaina, W. & Rashida, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Computer Science, 72, 414-422. doi: https://doi.org/10.1016/j.procs.2015.12.157. [GS Seach]

Souza, N., S.; Wives, L. K. & Perry, G. T. (2019). Tendências de Pesquisas que Utilizam Learning Analytics em MOOCs: um mapeamento sistemático. Renote: revista novas tecnologias na educação. v. 17, 1, p. 82-92. doi: https://doi.org/10.22456/1679-1916.95710

Sukhija, K.; Jindal, M. & Aggarwal, N. (2015) .The recent state of educational data mining: A survey and future visions. *Proceedings of the 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE).* Amritsar, India. doi: 10.1109/MITE.2015.7375344. [GS Search]

Sunar, A. S.; Abbasi, R. A; Davis, H. C.; White, S. & Aljohani, N. R. (2018). Modelling MOOC learners' social behaviours. *Jounal Computers in Human Behavior*, p. 1-12. doi: https://doi.org/10.1016/j.chb.2018.12.013. [GS Search]

Wagner, R.; Passerino, L.; Silveira, S.; Franciscatto, R. & Lima, J. V. (2016). SolAssist Learning: formação em tecnologias assistivas através de um MOOC e uma biblioteca virtual de soluções assistivas. *Revista Brasileira de Informática na Educação*. v. 24, no 3, p. 62-74. doi: http://dx.doi.org/10.5753/rbie.2016.24.3.62. [GS Search]

Waheeda, H.; Hassana, S. U.; Aljohanib, N. R.; Hardmand, J.; Alelyanic, S. & Nawazd, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Journal Computers in Human Behavior*, v. 104, p. 106-189. doi: https://doi.org/10.1016/j.chb.2019.106189. [GS Search]

Wen, Y.; Tian, Y.; Wen, B.; Zhou, Q.; Cai, G. & Liu, S. (2020). Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. Journal Tsinghua Science and Technology. v. 25, n. 3, p. 336-347. doi: 10.26599/TST.2019.9010013. [GS Search]

Wulf, J.; Blohm, I.; Leimeister, J. M. & Brenner, W. (2014). Massive open online courses. *Business & Information Systems Engineering.* v. 6, no. 2, p. 111–114. doi: https://doi.org/10.1007/s12599-014-0313-9. [GS Search]

Xing, W.; Chen, X.; Stein, J. & Marcinkowski, M. (2016). Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. *Jounal Computers in Human Behavior*. v. 58, p. 119-129. doi: https://doi.org/10.1016/j.chb.2015.12.007. [GS Search]