Revista EaD &
tecnologias digitais na educação

Proposta de Melhoria dos Dados de Relatórios de Uma Plataforma de MOOCs Brasileira: Um Estudo Introdutório Baseado em um Modelo para Mineração de Dados Educacionais

Vanessa Faria Souza

vanessa.souza@ibiruba.ifrs.edu.br

Instituto Federal de Educação Ciência e Tecnologia do Rio Grande do Sul.

Resumo: *Ambientes Virtuais de Aprendizado (AVAs) não foram projetados para mineração de dados. Como os dados que são originados do seu uso não são armazenados de maneira sistemática, sua análise completa requer pré-processamento longo e trabalhoso. A plataforma de Massive Open Online Courses (MOOCs), considerada para este estudo, não é díspar, muitos problemas são detectados na estruturação dos dados extraídos, quando se tenta realizar o processo de mineração de dados de um curso, na maioria dos casos os resultados alcançados não proporcionam conhecimentos relevantes com relação aos alunos. Diante deste fato, buscar princípios que forneçam mecanismos para melhoria nos dados armazenados na plataforma são válidos, pois auxiliam a melhorar o entendimento sobre os estudantes matriculados. Como estes aprendem, sobre suas preferências e anseios quanto aos cursos, sobre a probabilidade de conclusão e em especial sobre propensão a desistência. Assim, neste trabalho, é realizada uma análise dos dados gerados pela plataforma onde se descreve suas principais limitações, em seguida é apresentado um modelo de dados para facilitar a análise e mineração de dados educacionais, que representa uma abstração aperfeiçoada de quais elementos devem ser armazenados por um AVA para facilitar tal processo. Por fim após a avaliação deste modelo são feitas recomendações de aprimoramentos para a plataforma.*

Palavras-Chave: *Mineração de Dados Educacionais, Ambientes Virtuais de Aprendizagem, Modelo de Dados*

Abstract: *Virtual Learning Environments are not designed for data mining. As the data that originates from its use is not stored systematically, its complete analysis requires long and laborious pre-processing. The Massive Open Online Courses (MOOCs) plat-*

form, considered for this study, is not disparate, many problems are detected in the structure of the extracted data, when trying to perform the data mining process of a course, in most cases the results achieved do not provide relevant knowledge regarding students. Given this fact, looking for principles that provide mechanisms for improving the data stored on the platform are valid, as they help to improve the understanding of enrolled students. How they learn, about their preferences and desires for the courses, about the probability of completion and especially about the propensity to drop out. Thus, in this work, an analysis is performed of the data generated by the platform where its main limitations are described, then a data model is presented to facilitate the analysis and mining of educational data, which represents an improved abstraction of which elements should be stored by an AVA to facilitate such a process. Por fim após a avaliação deste modelo são feitas recomendações de aprimoramentos para a plataforma.

Keywords: *Educational Data Mining, Virtual Learning Environments, Data Model.*

1. Introdução

A chegada da Indústria 4.0 trouxe grandes mudanças para a sociedade e para o desenvolvimento científico, que afetam fortemente a maneira como as pessoas aprendem, ensinam e entendem o conhecimento e a educação. Neste sentido, a Educação a Distância (EAD) tem crescido em vários países, no Brasil por exemplo tem se estabelecido de forma abrangente. No censo realizado em 2018, constatou-se um aumento expressivo de 3,7 milhões para 7,7 milhões de matrículas, entre os anos de 2016 e 2017, bem como um aumento considerável entre 2017 e 2018 como pode ser visualizado na Figura 01 (ABED, 2019).

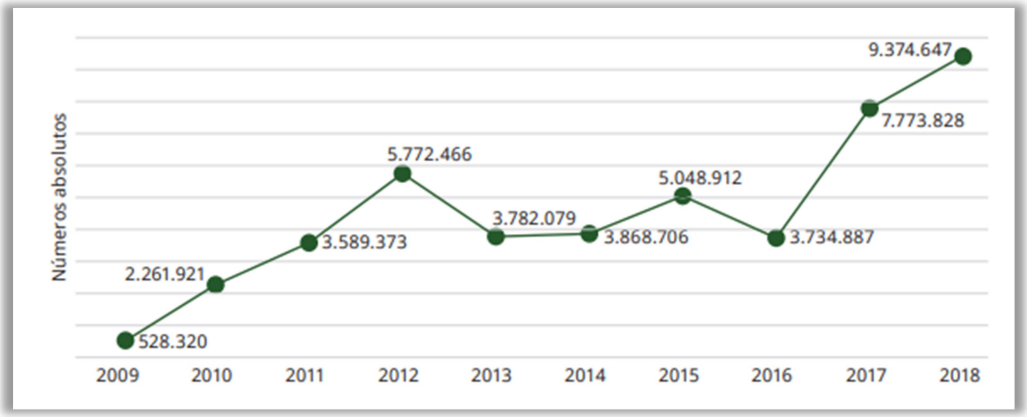

Figura 1: Evolução do total de matrículas contabilizadas pelo Censo EAD.BR
Fonte: ABED (2019)

Com o advento da EAD, há um número crescente de estudantes atraídos para o Massive Open Online Course (MOOC), com números expressivos de matrículas nos últimos anos (ROMERO & VENTURA, 2016). Em 2012, a edX, uma startup sem fins lucrativos de Harvard e do Massachusetts Institute of Technology, teve 370.000 alunos em seus primeiros cursos oficiais. O Coursera, fundado em janeiro de 2011, alcançou de 1,7 milhão de alunos registrados e está crescendo "mais rápido que o Facebook" (WANG, HU & ZHOU, 2018). Um curso de inteligência artificial, de Stanford oferecido em 2011, online e de forma gratuita e atraiu 160.000 estudantes (WANG, HU & ZHOU, 2018). Autores como Romero & Ventura (2016); Greene, Oswald & Pomerantz (2015); Hew, Qião & Tang

(2018); Wang, Hu & Zhou (2018) e Xing et al. (2015) salientam que o modelo MOOC de ensino-aprendizagem corresponde a um exemplo bastante aprimorado de educação sustentável. Além disso, ajuda a estabelecer uma educação personalizada, que é considerada como tendo grande potencial para promover também um desenvolvimento sustentável (WANG, HU & ZHOU, 2018).

Aponta-se como um grande diferencial dos MOOCs a grande quantidade de dados gerados pelas interações no Ambiente Virtual de Aprendizagem (AVA), o que abre novas possibilidades para estudar e compreender essas interações (ONAH, SINCLAIR & BOYATT, 2014). Desta forma, visualizou-se potencial para novos estudos e algumas áreas de pesquisa surgiram nos últimos anos com intuito de auxiliar em questões como essas. Por exemplo, a Mineração de dados Educacionais (MDE), que é uma área interdisciplinar que lida com o desenvolvimento de métodos para explorar dados originados no contexto educacional (ROMERO & VENTURA, 2016). A MDE é responsável pelo desenvolvimento de métodos para a extração de conhecimento a partir de base de dados educacionais, especialmente, de cursos online (KAMPFF, REATEGUI & LIMA, 2008; ROMERO & VENTURA, 2013). Durante o curso, é desejável que a MDE obtenha dados de interação dos estudantes e interprete os seus significados, a fim subsidiar professores em eventuais ajustes nas estratégias pedagógicas.

O crescente uso de AVAs, softwares educacionais e outras tecnologias que amparam o ensino por meio do computador, gera uma grande quantidade de dados, como já mencionado, no entanto, um grande problema, é que com tantas fontes de dados diferentes, existe uma falta de padronização na maneira como os dados são coletados e armazenados, ocasionando um grande esforço de pré-processamento de dados, que por si só, constitui um dos grandes desafios da MDE (COSTA et al, 2012). Segundo Krüeger, Merceron & Wolf (2010), é um fato bem conhecido que a compreensão e o pré-processamento de dados constituem os principais trabalhos do processo de análise e mineração de dados, os autores afirmam também que o AVA não foi projetado para análise e mineração dados. Mesmo que muitos softwares de aprendizado armazenem dados de uso, eles foram projetados para apoiar o aprendizado, não para analisar os dados ou mesmo para manter dados que possam ser minerados.

Contudo, o campo da MDE tem avançado precisamente porque valiosas informações são obtidas com a análise e mineração de dados armazenados pelo software educacional (ROMERO & VENTURA, 2007; BAKER & YACEF, 2009). Como consequência, o processo de MDE requer um longo pré-processamento (MACERON & YACEF, 2008). Ao corroborar essa afirmação, Souza & Perry (2019) salientam que dois dos maiores desafios enfrentados na MDE são a complexidade de manipulação dos dados gerados nos AVAs e a diversidade destes dados. Neste sentido, devido a observações realizadas em um Ambiente Virtual de Aprendizagem, que disponibiliza vários MOOCs, pode-se perceber que tanto a qualidade quanto a diversidade dos dados extraídos dessa plataforma, para realização da MDE, não são suficientes para obtenção de bons resultados.

Consequentemente, melhorias na forma de armazenar as interações entre usuário e a plataforma, assim como a inserção de mais atributos verificáveis na execução de um curso por um estudante, são fatores que seriam capazes de aumentar a qualidade e a diversidade dos dados gerados por ela. O que pode significar extração de conhecimentos mais reais e acurados via MDE. Nessa perspectiva, este artigo tem como objetivo realizar sugestões de melhorias no processo de armazenamento de AVAs, baseado em um modelo de dados proposto por Krüeger, Merceron & Wolf (2010).

2. Mineração de Dados Educacionais

De acordo com Costa et al. (2012) a área emergente de Mineração de Dados Educacionais procura desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo que se prestem a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem, tais como AVAs, Sistemas Tutores Inteligentes (STIs), entre outros. Com tais métodos visa-se, por exemplo, entender melhor o estudante no seu processo de aprendizagem, analisando-se sua interação com o ambiente (COSTA et al, 2012).

De maneira geral, o processo da Mineração de Dados (MD) está relacionado aos seguintes "pilares" gerais: (1) os dados (coleta e armazenamento), (2) a informação (dado analisado e com algum significado) e (3) o conhecimento (informação interpretada e aplicada). Os autores Han & Kamber (2016) destacam que a Mineração de Dados tem atraído muita atenção na indústria da informação e na sociedade como um todo nos últimos anos, devido à ampla disponibilidade de enormes quantidades de dados e à necessidade iminente de transformar esses dados em informação e conhecimento úteis. Existem diferentes metodologias para a aplicação da MD, mas por consenso de estudiosos (HAN & KAMBER, 2016) existe uma sequência básica de etapas, que compõem o processo para que este seja caracterizado como mineração de dados, e foram definidas a princípio por Fayyad, Piatetsky-Shapiro & Smyth, (1996) como descrito na Tabela 1:

Tabela 1: Fases da Mineração de Dados.

FASE	DESCRIÇÃO
SELEÇÃO	É onde se faz a seleção dos dados. Esta fase afeta diretamente a qualidade do resultado final, pois é onde se define quais dados, com suas possíveis variáveis (atributos) farão parte desta seleção. É uma fase muito complexa, visto que esses dados podem vir de fontes e estruturas diferentes (arquivos-texto, banco de dados, relatórios, logs de acesso, transações etc).
PRÉ-PROCESSAMENTO	É uma fase onde são tratados os dados que contêm algum tipo de problema, tais como os dados com valores ausentes ou discrepantes, e isso tem fator determinante na efetividade do algoritmo escolhido.
TRANSFORMAÇÃO	É a conversão dos dados em um formato comum, para a aplicação dos algoritmos. Se necessário, aqui pode se obter informações que faltam através da combinação ou transformação, que são os dados derivados, como, por exemplo, a idade de uma pessoa, que pode ser calculada através de sua data de nascimento.
MINERAÇÃO DE DADOS	São usadas várias estratégias para a visualização e diferentes técnicas de mineração, como a aplicação de algoritmos de Aprendizado de Máquina de classificação, agrupamento, regra de associação dentre outras.
INTERPRETAÇÃO / AVALIAÇÃO	São usadas variadas técnicas de interpretação e avaliação dos dados, isso depende do campo de pesquisa, mas todos com um intuito final: a informação.

A mineração de dados pode ser aplicada a diversos e diferentes domínios e contextos, tal como na Educação, utilizando-se de métodos de Aprendizado de Máquina (AM). Todavia, há a necessidade, por exemplo, de adequação dos algoritmos de mineração de dados existentes para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados. Por outro lado, há uma necessidade significativa e urgente no provimento de ambientes computacionais apropriados para mineração de dados educacionais, oferecendo facilidades de uso para cada um dos atores envolvidos, notadamente ao professor.

3. Trabalhos Relacionados

Pesquisadores que trabalham com MDE tem salientado em seus trabalhos a dificuldade na manipulação dos dados extraídos dos AVAs, tanto pela complexidade de manipulação dos dados gerados quanto pela diversidade. Neste ponto de vista, Hong, Wei & Yang (2017) utilizaram técnicas de MDE na predição do abandono em MOOCS, sua proposta se resumia em prever desistentes usando informações de atividades de aprendizagem dos alunos, eles aplicaram um classificador em cascata de duas camadas com uma combinação de três classificadores de aprendizado de máquina diferentes - Random Forest (RF), SVM (Support Vector Machine) e Logística Multinomial de Regressão (MLR) para previsão, tiveram resultados experimentais que indicam que a técnica é promissora na previsão de desistências atingindo 97% de precisão. Nesta mesma linha de pesquisa, Liang, Li & Zheng (2016) mineraram dados de 39 cursos da plataforma XuetangX, essa plataforma possui código aberto é baseada na Edx. Os autores utilizaram uma abordagem de classificação supervisionada, usando como atributos o comportamento dos usuários e dados de log e alcançaram 89% de acurácia na predição de desistência em Cursos MOOC com algoritmo de árvore de decisão.

Embora, estas pesquisas tenham obtido sucesso no processo MDE desenvolvido, os autores deixam claro em suas análises que um dos principais entraves para aplicação de algoritmos de aprendizagem de máquina, vinculados a mineração de dados em MOOCs, são os dados disponíveis, que precisam ser trabalhados (pré-processamento e transformação) antes de serem utilizados no treinamento dos algoritmos e no processo de predição. Cabe salientar que no trabalho de Liang, Li & Zheng (2016), estes descrevem os detalhes da seleção de dados da plataforma, pré-processamento de dados, engenharia de recursos e teste de desempenho nos modelos de classificação que utilizaram, desta forma fica claro todo o esforço necessário para conseguir produzir bons resultados. Como também, Xing et al. (2016) realizaram uma predicação temporal de desistências em MOOCs, baseado nos algoritmos de Rede Bayesiana Geral (GBN), e em árvore de decisão (C4.5). Os dados analisados foram referentes a participação em fóruns, chats, e atividades realizadas. Esse trabalho apresenta como resultados uma abordagem de generalização de agrupamentos para construção de previsões mais robustas e precisas, para lidar com a variabilidade de dados apresentados pelos MOOCs.

Nessa perspectiva, Ruipérez-Valiente et al. (2015), que em seu trabalho descrevem o software ALAS-KA, que fornece um suporte a análise de dados gerados pela plataforma Khan Academy. O ALAS-KA fornece diferentes tipos de visualizações de informações, que não estavam disponíveis anteriormente na plataforma Khan Academy, inclui visualizações de informações individuais e da turma que podem ser utilizados para verificar os estilos de aprendizagem dos alunos com base em todos os indicadores disponíveis. O software que aplica a MDE auxilia professores e alunos a tomar decisões no processo de aprendizagem. Contudo, os bons resultados obtidos pelas pesquisas elaboradas por Xing et al. (2016) e Ruipérez-Valiente et al. (2015), estes salientam que tratar da grande diversidade de ações estudantis que podem ser capturadas pelo AVA, desde a escrita de uma pergunta em um Chat, até a visualização de um vídeo, é uma tarefa complexa. Nessa perspectiva afirmam ser complicada a forma de lidar com o fator da variabilidade de dados apresentados pelos MOOCs. Portanto, nota-se que a manipulação dos dados extraídos das plataformas de ofertas de MOOCs, é um fator que necessita de atenção de pesquisadores da área.

4. Análise dos Dados Extraídos de uma Plataforma de MOOCs Brasileira

A plataforma que foi analisada neste trabalho é uma instalação do Moodle, com um tema customizado. O formato empregado nos cursos segue um modelo padrão: o conteúdo é transmitido prioritariamente na forma de vídeos, mas também são usados textos, imagens e outros materiais que possam ser inseridos no Moodle. Na plataforma todos os MOOCs têm um vídeo de apresentação que fica disponível, mesmo sem o cadastramento do participante; os cursos possuem blocos com informações sobre o curso e sobre os professores, e as avaliações se dão na forma de testes de múltipla escolha (com a atividade "questionário", do Moodle). Em janeiro de 2021, havia 50 cursos disponíveis, com mais de 50.000 usuários cadastrados.

Os MOOCs disponibilizados na plataforma são auto formativos e não existe interação com professores ou tutores. Desta forma, qualquer ferramenta disponível no Moodle, que não exija obrigatoriamente a presença de um professor ou tutor acompanhando o curso, pode ser utilizada. Um exemplo disto são os fóruns de discussão, ferramenta do Moodle que, em alguns cursos da plataforma, possuem propostas bem direcionadas pelos professores/ autores, entretanto, os professores na maioria dos cursos não interagem com os alunos.

Com relação aos dados aos quais se podem aplicar os métodos de mineração de dados, estes são restritos. Cabe salientar que técnicas para mineração de texto, também são possíveis de serem aplicadas nas interações dos usuários nos Fóruns, que estão presentes em todos os cursos, e contêm muitas informações relevantes, mas esse não é o foco deste trabalho.

O MOOC do qual os dados são usados como exemplo neste artigo pertence a área de Humanas e teve 850 alunos matriculados. Com taxa de conclusão, na primeira edição de 49%. Ao se inscrever no curso primeiramente o aluno deve responder um questionário de perfil onde o aluno inclui informações como: nome, idade, formação. Este também possui questões relacionadas ao curso como: "se o aluno já havia praticado esportes ao ar livre" ou "se possui conhecimento teórico a esse respeito". Esse questionário é importante para responder questionamentos como: (1) O no nível ou área de formação do aluno influencia na conclusão do MOOC; (2) A razão descrita pela qual o aluno se interessou pelo MOOC tem relação com a desistência dos alunos. Desta forma, é importante manter este tipo de interação.

Mas o foco dessa pesquisa são as planilhas geradas a partir das realizações de atividades dos alunos. As atividades que podem ser realizadas nesse curso são, assistir os vídeos disponibilizados pelo professor (os vídeos se dividem em disponíveis no Youtube e vídeo aulas gravadas pelo professor), realizar a leitura das referências, participar dos fóruns e responder aos questionários de múltipla escolha. Cabe salientar que o curso estava dividido em 4 módulos e todos possuíam tais atividades para os alunos, exceto o ultimo que inclui apenas o fórum e a leitura das referências, e para finalizar o aluno deveria responder uma avaliação geral sobre o curso.

As planilhas geradas correspondem aos questionários de múltipla escolha respondidos pelos alunos, esses dados são insuficientes para minerar e chegar a conhecimentos relevantes, na Tabela 2 pode ser visualizado uma parte de uma planilha extraída do primeiro questionário do curso analisado. Os nomes dos alunos foram substituídos por números. A coluna um armazena a informação sobre o sobrenome do aluno, na co-

luna dois o nome, na coluna três e quatro a instituição e departamento onde o aluno está lotado. A seguir na coluna cinco o endereço de e-mail do aluno, depois na coluna seis o estado da atividade se foi finalizada, ou está em progresso, na coluna sete temos a data de início da atividade, depois na oito a data de finalização e na nove o tempo gasto para realização. A seguir na coluna dez está a nota final do aluno que é um somatório das colunas onze a quinze, que são as notas individuais do aluno em cada questão que compõe a atividade.

Tabela 2: Parte de planilha gerada pela Plataforma.

Sobrenome	Nome	Instituição	Departamento	Endereço de email	Estado	Iniciado em	Completo	Tempo utilizado	Avaliar/ 10,00	Q. 1/ 2,00	Q. 2/ 2,00	Q. 3/ 2,00	Q. 4/ 2,00	Q. 5/ 2,00
01	01			l01@hotmail.com	Finalizada	16 abril 2018	16 abril 2018	1 minuto 23 segundos	8,00	2,00	2,00	1,33	2,00	0,67
02	02			02@yahoo.com.br	Finalizada	4 maio 2018	4 maio 2018	4 minutos 32 segundos	8,67	2,00	2,00	1,33	2,00	1,33
03	03			03@yahoo.com.br	Em progress	4 maio 2018	-	-	-	-	-	-	-	-
04	04			04@gmail.com	Finalizada	5 maio 2018	5 maio 2018	3 minutos 18 segundos	8,67	2,00	1,33	2,00	2,00	1,33
05	05			05@yahoo.com.br	Finalizada	5 maio 2018	5 maio 2018	4 minutos 4 segundos	8,67	2,00	0,67	2,00	2,00	2,00
06	06			06@gmail.com	Finalizada	5 maio 2018	5 maio 2018	8 minutos 52 segundos	9,33	2,00	2,00	1,33	2,00	2,00

Fonte: Autores

5. Descrição do Modelo de Dados Krüeger, Merceron & Wolf (2010)

Cabe ressaltar que este modelo foi selecionado para essa pesquisa, embora já faça alguns anos que foi publicado, porque nas principais bases de dados pesquisadas: IEEExplore, ACM Digital Library, o portal de periódicos da Capes, SciELO e Springer; não foi encontrada outra publicação mais recente de um modelo de dados especifico para AVAs, para facilitar a MDE.

Os autores Krüeger, Merceron & Wolf (2010) definem que a funcionalidade de um AVA pode ser dividida em três partes principais: Gerenciamento de recursos de aprendizado, gerenciamento de usuários e comunicação entre usuários. No entanto, estatísticas e relatórios geralmente são básicos. Para ilustrar esse último ponto, os autores listam algumas perguntas que os relatórios de instalações da maioria dos AVAs utilizados não podem lidar: (1) Quantos alunos nunca viram o Recurso de Aprendizagem A? (2) Se os alunos se saem bem na atividade B, eles também se saem bem na atividade C? (3) Se os alunos resolverem o exercício D, eles também resolverão o exercício E? (4) Qual é a nota média no questionário G obtida pelos alunos que visualizaram o recurso F? (5) Quais cursos usam muitos recursos de aprendizado de áudio e/ou vídeo?

Para gerenciar adequadamente os alunos, especialmente os alunos a distância, é muito importante obter uma boa visão geral de seus comportamentos de aprendizagem. O objetivo do modelo de dados proposto Krüeger, Merceron & Wolf (2010) é complementar o AVA em todos os aspectos relacionados à análise e mineração de dados. O modelo proposto preocupa-se em descrever e estruturar as interações dos usuários com objetos de aprendizado armazenados no AVA. O modelo foi pensado para ser aplicado junto a plataforma Moodle.

6. O Modelo de Dados

O modelo de dados contém três tipos de tabelas: tabelas para descrever objetos encontrados no AVA, essas tabelas podem ser vistas como tabelas de dimensões; tabelas para descrever interações com objetos de aprendizagem, essas tabelas podem ser vistos como tabelas de fatos; e a terceira, tabelas de associação para descrever associações entre objetos.

Segundo Krüeger, Merceron & Wolf (2010) a escolha de ter uma tabela por objeto vem da observação de que a maioria dos AVAs possui um conjunto limitado de objetos que os professores conseguem manipular, a adoção desse mesmo conjunto deve facilitar para que eles possam analisar o uso de recursos pelos alunos.

6.1. O Esquema

O Modelo proposto traz algumas suposições gerais: (1) O AVA contém usuários e cursos. Os usuários podem se inscrever ou entrar em cursos e sair de cursos. Os usuários podem ter funções como "professor", "administrador", "tutor", "aluno"; (2) Um usuário pode ter diferentes funções em diferentes cursos. Por exemplo, um usuário pode ser um tutor no curso "Introdução à programação" e ser aluno no curso "História americana"; (3) Um AVA pode conter grupos associados a cursos. Alunos se matriculam nesses cursos; (4) Um AVA contém fóruns, wikis, recursos (como vídeos e áudios) e questionários. Neste modelo o questionário é qualquer tipo de tarefa, exercício ou teste que um professor possa desejar aplicar aos alunos; (5) Fóruns, wikis, recursos e questionários estão associados a cursos. Assim, um recurso por exemplo, pode ser usado em vários cursos; e (6) Um questionário pode conter uma ou mais perguntas que também estão contidos no AVA. As perguntas estão associadas a questionários, e a uma determinada pergunta pode ser associada a vários testes.

Essas premissas gerais abrangem o caso particular do AVA onde recursos, fóruns, wikis ou questionários existem apenas dentro de um determinado curso. Nesse caso específico, uma tabela de associação contém apenas uma tupla. Assumimos que um AVA registra ou armazena interações de usuários. Para qualquer interação, este armazena a identificação do usuário, do recurso, fórum, wiki, questionário, bem como a timestamp (horário real), a natureza da interação (“visualização”, “modificação”, “criação”, “tentativa”, “envio”, dentre outros), as marcas e a contribuição, quando relevante. A Figura 03 contempla as tabelas que são fundamentais no modelo de dados proposto.

ISSN 2318-4051

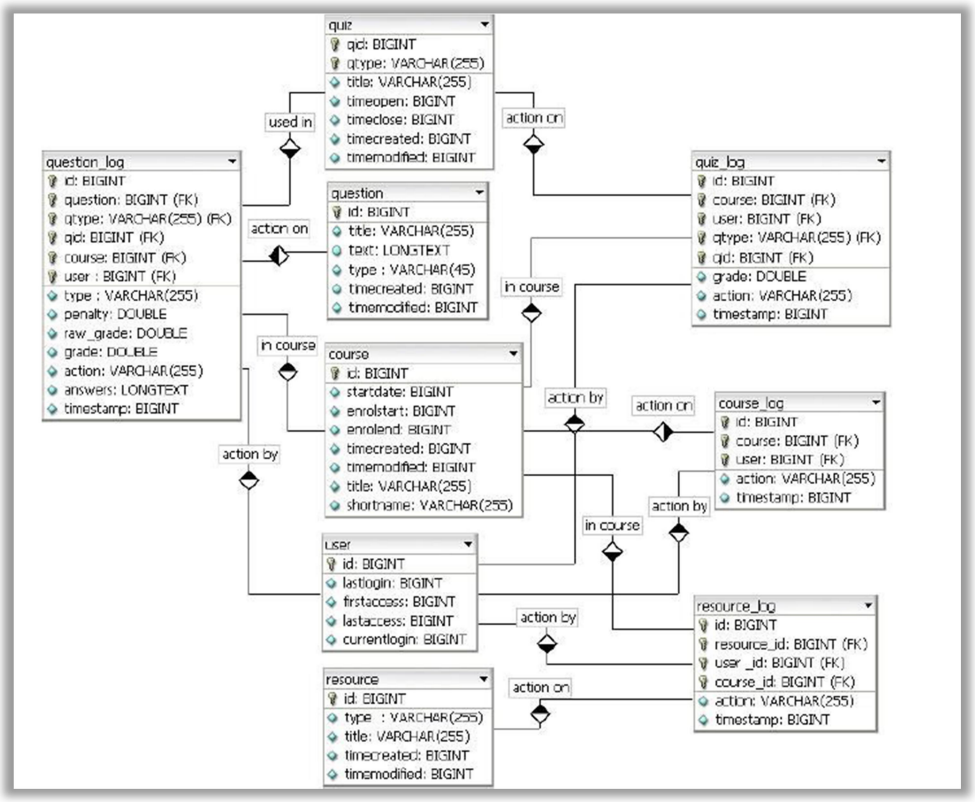

Figura 3: Esquema Relacional do Modelo
Fonte: Adaptado de Krüeger, Merceron & Wolf (2010

Cada tabela contém um ID do elemento, que é a chave ou o identificador da tupla. As cinco tabelas centrais descrevem objetos geralmente encontrados no AVA, suas caracterizações podem ser visualizadas na Tabela 3.

Tabela 3: Descrição das Tabelas de objetos do Modelo.

TABELA	DESCRIÇÃO	ELEMENTOS
Usuário	Descreve os usuários registrados no AVA.	Os fistaccess e lastaccess são as horas e datas em que um usuário acessou pela primeira vez e pela última vez qualquer tipo de objeto de aprendizagem, como um recurso ou um questionário. Lastlogin e currentlogin são os horários e as datas em que um usuário efetuou login no AVA pela última vez, respectivamente. Observe que um usuário pode efetuar login sem acessar nenhum objeto.
Curso	Descreve os cursos existentes no AVA. Um curso existe por um determinado período de tempo.	Startdate é a hora e a data que o curso deve começar, esse horário geralmente é fixado pelo professor responsável. O elemento enrolstart é a data em que os usuários podem se inscrever neste curso e o elemento emrolend é a data em que os usuários não podem mais se inscrever no curso. O elemento timemodified é a hora e a data em que este curso foi modificado pela última vez. O title e os elementos e shortname são autoexplicativos.
Quiz	Descreve os testes existentes no AVA.	Qtype é o tipo do questionário. Pode aceitar valores como "atribuição", "pontuação" e assim por diante, de acordo com os diferentes tipos de questionários disponibilizados. O qid combinado com o tipo compõe a identificação de um questionário. O title é o título que o professor responsável atribui para este questionário. Os elementos timeopen e timeclose se referem às datas e horários disponibilizados para os alunos responderem, enquanto timecreated é a data e a hora em que o questionário foi criado, timemodified é a data e hora em que o questionário foi respondido a última vez que foi modifica-

Revista: EaD & Tecnologias Digitais na Educação, Dourados, MS, 2021 – n° 11, Vol. 9
ISSN 2318-4051

		do.
Questionário	Descreve as perguntas que compõem os questionários.	Title é o título da pergunta, enquanto o text é o texto do problema a ser resolvido. O type é uma categoria como "múltipla escolha", "verdadeiro-falso". Timecreated e timemodified tem as mesmas definições da tabela Quiz.
Recurso	Descreve os recursos disponíveis.	Type descreve o tipo de recurso como "arquivo", "url", “Diretório”, “áudio”, “imagem” e assim por diante. timecreated e timemodified tem as mesmas definições da tabela Quiz.. Title é o título deste recurso.

Fonte: Adaptado de Krüeger, Merceron & Wolf (2010

As tabelas restantes do modelo descrevem interações com objetos de aprendizagem. São os itens que são armazenados enquanto os usuários usam objetos de um AVA. Estas são detalhadas na Tabela 4.

Tabela 4: Descrição das Tabelas de interações do Modelo.

TABELA	DESCRIÇÃO	ELEMENTOS
Quis_log	Descreve as informações que um AVA deve armazenar quando um usuário interage com questionários	ID (é a chave) do usuário que interagiu. O course é o ID (a chave) do curso em que a interação ocorreu. Os qid e qtype referem-se ao questionário que foi abordado. Grade é a pontuação obtida no questionário, timestamp fornece a data e hora da interação. O elemento action fornece o tipo de ação que ocorreu. Uma ação pode ser "visualizar", nesse caso, o usuário simplesmente olhou para o questionário, "tentativa"; nesse caso, o usuário tentou fazer questionário, "enviar"; nesse caso o usuário tentou e terminou o questionário, "modificar" se o questionário foi modificado dentre outros.
Question_log	Descreve as informações que um AVA deve armazenar quando um usuário interage com uma pergunta de um questionário.	Contém todos os elementos já incluídos na tabela quiz_log. O penalty fornece as marcas de penalidade dadas nessa interação. Se um questionário for executado no modo adaptativo, o aluno poderá tentar novamente a pergunta após um erro de resposta. Nesse caso, pode-se querer impor uma penalidade para cada resposta errada a ser subtraído da nota final da pergunta. A quantidade de penalidade é escolhida individualmente para cada pergunta ao configurar ou editar a pergunta. O raw_grade fornece a pontuação bruta obtida nessa interação. O grade fornece a pontuação nessa pergunta e interação quando a penalidade é levada em consideração. ID identifica a pergunta que foi abordada, type pode aceitar valores como "múltipla escolha", "verdadeiro / falso" e answer ou answers, pode aceitar uma ou várias respostas, dadas pelo usuário na interação.
Resource_log	Descreve as informações que um AVA deve armazenar quando um usuário interage com recursos.	Contém elementos semelhantes aos da tabela quiz_log.

Fonte: Adaptado de Krüeger, Merceron & Wolf (2010

Neste artigo, foram apresentados apenas os aspectos principais do modelo, que podem auxiliar nas melhorias previstas para a plataforma a qual se refere essa investigação, são omitidas as tabelas de associação que constam no respectivo modelo para relacionar os objetos um ao outro.

7. Recomendações para o Armazenamento de Dados e Geração de Relatórios

Com base na análise feita sobre o modelo de dados proposto por Krüeger, Merceron & Wolf (2010), identifica-se que algumas adequações, possíveis de serem implementadas no AVA analisado, que podem melhorar significativamente a quantidade de atributos armazenados, que atualmente é baixa, assim como melhorar a qualidade dos dados que são mantidos pela plataforma, em consequência sua geração de relatórios. Elementos, esses, que colaboram no processo de MDE, que pode auxiliar na descoberta de novos conhecimentos sobre o público que se utiliza dessa plataforma, assim melhorando a experiência desses alunos. Na Tabela 5 estão sistematizadas as principais recomendações de ajustes no armazenamento de dados proposto.

Tabela 5: Sugestões de Adequação.

ADEQUAÇÕES		DESCRIÇÃO
ARMAZENAR	Interações com recursos	A plataforma deve ser capaz de armazenar dados que dizem respeito a ações do usuário em todos os recursos disponibilizados. Desta forma, teria como saber se ele assistiu um vídeo completo, ou até que ponto ele assistiu, se ele ouviu um áudio completo ou em quanto tempo ele parou de ouvir.
	Mais informações nas atividades de questionário	Todos o tipo de ação efetuada por um usuário em um questionário deve ser armazenada, as ações podem ser: "visualizar"; "tentativa"; "enviar"; "modificar" como descrito no modelo de dados.
	Interações com páginas de conteúdos	A plataforma deve armazenar dados quanto a rolagem da página pelo usuário, deve-se saber se este rolou a página até o final, ou seja, se visualizou todo o conteúdo disponibilizado.
INSERIR	Questionário de perfil com mais questões sobre a temática do curso	Mais questões sobre a afinidade do usuário com curso.
	Itens de interatividade em recursos que os alunos possam responder	Inserir perguntas simples no decorrer de um vídeo ou outro recurso para que o usuário tenha mais interatividade com a plataforma, e armazenar as respostas a estes questionamentos.
	Questionário de opinião após recursos	Depois da exibição de um recurso inserir questões curtas sobre aquele material, não especifico sobre o conteúdo, mas sobre a qualidade do recurso.

Fonte: Autora

Como pode ser observado na Tabela 05 o principal obstáculo para MDE no AVA investigado é que a plataforma não armazena informações de todas as ações que o usuário pode executar. Apenas os questionários tem dados mantidos no sistema, referentes a notas e estado. Desta forma não é possível verificar o quanto o aluno está envolvido com o curso, se ele realmente lê os materiais, assiste as vídeo aulas, que são elementos centrais para se obter mais informações sobre estes usuários. Esse é um levantamento preliminar, mas com essas adequações e aplicação do processo de MDE em mais cursos, mais itens podem ser pensados para implementações futuras.

8. Considerações Finais

Como citado por Krüeger, Merceron & Wolf (2010), o AVA não foi projetado para análise e mineração de dados, eles armazenam dados, mas não foram projetados com essa finalidade, e sim para apoiar o ensino e o aprendizado. Todavia a MDE tem avançado justamente devido aos dados obtidos com mineração destas plataformas. Recentes pesquisas (SOUZA & PERRY, 2019) apontam que alguns dos maiores desafios enfrentados na MDE são vinculados a qualidade dos dados obtidos nas plataformas. Nesta perspectiva, o modelo de dados desenvolvido por Krüeger, Merceron & Wolf (2010) auxilia na detecção de itens que necessitam de adequação nos AVAS para que possam fornecer dados apropriados para um bom processo de mineração.

O padrão proposto pelos autores sistematiza um modelo relacional de dados que especifica quais são os principais elementos que devem fazer parte das tabelas mantidas por um AVA. Além dos itens comuns como questionários, usuários, cursos, chama atenção no modelo é que este propõe armazenar todas as interações que o usuário experienciar com a plataforma, visualizações de recursos, ações referentes a entrega de uma atividade e até mesmo interações sem acesso a atividades ou recursos, e que estas gerem dados que possam ser analisados posteriormente. Essa gama ampliada de dados armazenados, preconizada pelo modelo confere ao processo de MDE mais hipóteses para testar com os algoritmos de aprendizagem de máquina, o que possibilita gerar conhecimentos mais acurados acerca dos alunos matriculados nos MOOCs.

No caso particular deste trabalho, que trata de um ambiente baseado no Moodle, foi demostrado que os dados extraídos para realização da MDE não são suficientes para alcançar maior entendimento sobre os alunos matriculados em seus cursos, desta forma fica difícil perceber quais motivos levam os alunos a permanecerem ou evadirem um curso. Em termos de gerenciamento da plataforma isto é um grande desafio.

Melhorias na forma de armazenamento dos dados e na quantidade dos atributos mantidos, são fatores que podem melhorar a qualidade dos dados gerados pela plataforma, neste sentido com base no modelo que se tomou como referência é possível destacar que o AVA precisa implementar algumas adequações, entre elas destaca-se: Armazenamento de interações com recursos, armazenamento de interação com páginas de conteúdo e armazenamento de mais informações nas atividades de questionário.

A inserção destes atributos aumenta o conjunto de dados que podem ser minerados e consequentemente melhora a compreensão sobre como os alunos aprendem, quais fatores influenciam em sua permanência no curso e principalmente ajuda na percepção de quais alunos são mais propensos a evadir. Se os algoritmos de MDE forem bem treinados, logo nas primeiras semanas do lançamento de um curso é possível analisar os dados de alunos e identificar aqueles inclinados a desistir e ações individuais podem ser implementadas, já que promover essas ações para todos os alunos de um MOOC é inviável.

Como trabalhos futuros pretende-se analisar a arquitetura de implementação do modelo de Krüeger, Merceron & Wolf (2010) e implementá-lo na plataforma foco desta pesquisa, pois de acordo com os autores esse modelo foi elaborado para ser implantado junto a plataforma Moodle e como o AVA em questão é uma instalação do Moodle, há a possibilidade de realizar essa implementação, a qual será posteriormente descrita no formato de artigo, para auxiliar mais pessoas que pretendam melhorar a qualidade dos dados gerados por ambientes como esse.

Referências

ABED - CENSO EAD. 2019 [Online]. Relatório analítico da aprendizagem a distância no Brasil. Censo EAD BR 2017. Disponível em: <http://www.abed.org.br>. Acesso em 21 de março de 2020.

BAKER, S. J. D. R; & YACEF, Y. The State of Educational Data Mining in 2009: A Review and Future Visions. (2009). JMDE - Journal of Educational Data Mining, 1(1), p. 3-17. doi: https://doi.org/10.5281/zenodo.3554657 [GS Search]

COSTA, E.; BAKER, R. S. J.; AMORIM, L.; MAGALHÃES, J. & MARINHO, T. (2012). Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. In: Jornada de Atualização em Informática na Educação (JAIE), Rio de Janeiro, 1, p. 1-29. [GS Search]

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; & SMYTH, E P. (1996). From data mining to knowledge discovery: An overview. In: Advances in knowledge discovery and data mining. AI Magazine, 17(3), p. 1–34. doi: https://doi.org/10.1609/aimag.v17i3.1230 [GS Search]

GREENE, J. A.; OSWALD, C. A.; & POMERANTZ, J. (2015). Predictors of Retention and Achievement in a Massive Open Online Course. American Educational Research Journal, 52(5), p. 925–955. doi: https://doi.org/10.3102/0002831215584621 [GS Search]

HAN, J.; & KAMBER, E M. (2006). Data mining: Concepts and techniques. 2ed, 500 Sansome Street, Suite 400, San Francisco, CA 94111: Morgan Kaufmann Publisher. [GS Search]

HEW, K. F.; QIAO, C.; & TANG, Y. (2018). Understanding Student Engagement in Large-Scale Open Online Courses: A Machine Learning Facilitated Analysis of Student's Reflections. In 18 Highly Rated MOOCs. International Review of Research in Open and Distributed Learning, 19(3), p. 69-93. doi: https://doi.org/10.19173/irrodl.v19i3.3596 [GS Search]

HONG, B.; WEI, Z.; & YANG, Y. (2017). Discovering Learning Behavior Patterns to Predict Dropout in MOOC. In 12th International Conference on Computer Science and Education (ICCSE), Houston, TX, USA, p. 700–704. doi: 10.1109/ICCSE.2017.8085583 [GS Search]

KAMPFF, A.; REATEGUI, E.; & DE LIMA, J. (2008). Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. RENOTE, 6(1), p. 1-8. doi: https://doi.org/10.22456/1679-1916.14394 [GS Search]

KRÜEGER, A.; MERCERON, A.; & WOLF, B. (2010). A Data Model to Ease Analysis and Mining of Educational Data. In: International Conference on Educational Data Mining, (MDE), 3, p. 131-140, Pittsburgh, PA, USA. [GS Search]

LIANG, J.; LI, C.; & ZHENG, L. (2016). Machine Learning Application in MOOCs: Dropout Prediction. In 11th International Conference on Computer Science & Education (ICCSE 2016), Nagoya University, Japan, p. 752–57. doi: 10.1109/ICCSE.2016.7581554 [GS Search]

MERCERON, A.; & YACEF, K. (2014). Interestingness Measures for Association Rules in Educational Data. In: First International Conference on Educational Data Mining, p. 57-66. [GS Search]

ONAH, D. F.; SINCLAIR, J.; & BOYATT, R. (2014). Dropout rates of massive open online courses: behavioural patterns. In 14th, EDULEARN, EUA, p. 5825-5834. [GS Search]

ROMERO, C.; & VENTURA, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, p. 125-146. doi: https://doi.org/10.1016/j.eswa.2006.04.005 [GS Search]

ROMERO, C.; & VENTURA, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), p. 12-27. doi: **https://doi.org/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, 7(1). doi: **https://doi.org/10.1002/widm.1187** [GS Search]

RUIPÉREZ-VALIENTE, J. A.; MUÑOZ-MERINO, P. J.; LEONY, D.; & KLOOS, C. D. (2015). Alas-Ka: A learning analytics extension for better understanding the learning process in the Khan Academy platform. Computers in Human Behavior, 47, p. 139-148. doi: https://doi.org/10.1016/j.chb.2014.07.002 [GS Search]

SOUZA, V. F.; & PERRY, G. (2019). Identifying student behavior in MOOCs using Machine Learning. International Journal of Innovation Education and Research, 7(3), p. 30-39. [GS Search]

WANG, L.; HU, G.; & ZHOU, T. (2018). Semantic Analysis of Learners Emotional Tendencies on Online MOOC Education. Sustainability, 10(6). doi: **https://doi.org/10.3390/su10061921** [GS Search]

XING, W.; WADHOLM, R.; PETAKOVIC, E.; & GOGGINS, S. (2015). Group learning assessment: developing a theory-informed analytics. Journal of Educational Technology & Society, 18(2), p. 110-128. [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. Elservier, Computers in Human Behavior, 58, p. 119-129. doi: https://doi.org/10.1016/j.chb.2015.12.007 [GS Search]

Agradecimentos: O presente trabalho foi realizado com apoio do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS).