

Mineração de dados educacionais na análise das interações dos alunos em um Ambiente Virtual de Aprendizagem

Ricardo E. D. e Silva¹, Jorge L. C. Ramos¹, Rodrigo L. Rodrigues²,
Alex Sandro Gomes³, José Alexandro V. Fonsêca⁴

¹Colegiado de Engenharia da Computação – Universidade Federal do Vale do São Francisco - UNIVASF

²Departamento de Educação – Universidade Federal Rural de Pernambuco – UFRPE

³Centro de Informática – Universidade Federal de Pernambuco – UFPE

⁴Núcleo de Educação a Distância – Universidade de Pernambuco – UPE

jorge.cavalcanti@univasf.edu.br, riceuller@gmail.com, rlr@ded.ufrpe.br,
asg@cin.ufpe.br, alexandrovf@gmail.com.

Abstract. *This paper aims to describe the analysis of data from database of the virtual environment Moodle, from Federal University of São Francisco Valley - Brazil (UNIVASF), using educational data mining from the cluster analysis techniques. Specifically, were analyzed the records of students' interactions in the virtual environment promoting the formation of different groups of students with similar interaction features. After analyzing these patterns obtained were used to determine the knowledge of the student groups, enabling insight into their interactions and their performance in the course.*

Resumo. *Este artigo tem como objetivo descrever as análises dos dados oriundos do banco de dados do ambiente virtual Moodle da Universidade Federal do Vale do São Francisco (UNIVASF), utilizando a mineração dos dados educacionais, a partir da técnica de análise de agrupamentos. Foram analisados especificamente os dados dos registros das interações dos alunos no ambiente virtual promovendo a formação de grupos distintos de alunos com características de interação semelhantes. Depois de analisados, esses padrões obtidos foram utilizados na determinação do conhecimento acerca dos grupos de estudantes, possibilitando uma percepção sobre suas interações e respectivo desempenho na disciplina.*

1. Introdução

Com crescimento acentuado nos últimos anos, a Educação a Distância (EAD) tem assumido um importante papel no cenário educacional, em particular na formação em nível superior de setores da população que dificilmente alcançariam essa categoria de escolaridade, por fatores sociais e econômicos diversos. Essa modalidade viabiliza a formação em horário flexível distinto dos horários em ambientes escolares presenciais, possui um alcance geográfico muito abrangente, assim como possibilita redução de custos na formação profissional do cidadão.

Dentre as diversas modalidades da EAD, destacamos o *e-Learning*, que possui como foco principal a autoaprendizagem, com a mediação de recursos didáticos sistematicamente organizados, apresentados em diferentes suportes tecnológicos de informação, utilizados isoladamente ou combinados, e veiculados através da Internet. Esta modalidade possibilitou a utilização dos chamados Ambientes Virtuais de Aprendizagem (AVAs).

Esses ambientes facilitam a interação entre alunos, professores e o conteúdo oferecido pelos cursos, onde são disponibilizados aos usuários o acesso a recursos e ferramentas que permitem a comunicação, construção do conhecimento, avaliação e ainda, um melhor acompanhamento

da interação dos usuários, pois permitem o armazenamento, em seu banco de dados, de todas as interações e informações do ambiente, gerando uma grande quantidade de dados que podem vir a gerar conhecimento aplicável a respeito dos usuários, disciplinas e cursos.

A fim de construir conhecimento a partir desses dados, podem ser utilizadas técnicas de Mineração de Dados Educacionais, do inglês: *Educational Data Mining* (EDM). A EDM é definida como a área responsável pelo o desenvolvimento de métodos de extração de conhecimento a partir de bases de dados de ambientes educacionais. [Baker et al. 2011]

Com este trabalho, buscou-se coletar e analisar os dados das interações em um ambiente virtual de aprendizagem, de alunos iniciantes na educação a distância, a partir da utilização de uma técnica da mineração de dados educacionais, possibilitando a descoberta de conhecimento relevante acerca desses alunos.

Este artigo contempla mais quatro seções. Na seção seguinte, é apresentado o embasamento teórico e análise de trabalhos relacionados para subsidiar as análises e discussões apresentadas. Na seção três, define-se o percurso metodológico, com a descrição do experimento, seus procedimentos, a coleta e o tratamento dos dados. Na quarta seção são apresentados os resultados do estudo, com as respectivas discussões. Por fim, na seção cinco são apresentadas as considerações finais e sugestões de trabalhos futuros.

2. Fundamentação Teórica

Já não se discute mais a importância do uso das Tecnologias de Informação e Comunicação (TIC) na educação; seja ela em qualquer modalidade, nível ou dispositivo. Nas novas áreas de pesquisa sobre a associação das TIC com a educação, o que se busca é encontrar e apresentar novos métodos, técnicas e procedimentos que: aperfeiçoem essa relação, facilitem a aprendizagem, permitam um melhor planejamento e gestão de cursos e ofereçam alternativas para superar desafios e dificuldades em estruturas educacionais diversas [Rodrigues et al. 2014].

2.1. e-Learning e Ambientes Virtuais de Aprendizagem

O *e-Learning* refere-se ao uso de meios eletrônicos e tecnologias de informação e comunicação (TICs) na educação [Clark and Mayer 2011]. O *e-Learning* pode proporcionar uma aprendizagem personalizada, em conformidade com a necessidade, disponibilidade e o ritmo do indivíduo, independentemente da plataforma usada para conectá-lo à Internet. É uma evolução necessária no contexto educativo, face aos requisitos da sociedade atual: uma sociedade da informação, de aprendizagem e do conhecimento [Gonçalves 2007].

Segundo [Anderson 2008], as interações que acontecem em um ambiente de *e-learning* são: i) aluno-aluno, ii) aluno-professor, iii) aluno-conteúdo, iv) professor-conteúdo, v) professor-professor e vi) conteúdo-conteúdo. Dadas as características do curso analisado e do experimento realizado neste trabalho, foram coletadas e analisadas as interações i, ii e iii, considerando a interação aluno-tutor como sendo uma variação da interação aluno-professor.

Já os AVAs são sistemas que sintetizam as funcionalidades de softwares para Comunicação Mediada por Computador (CMC) e os métodos de entrega de material de cursos online. Muitos desses sistemas reproduzem a sala de aula presencial física para o meio online, outros buscam, além de simplesmente reproduzir ambientes educacionais existentes para um novo meio, fazer uso da tecnologia para propiciar a constituição de novos ambientes educacionais, de acordo com a natureza específica desses diferentes meios [Schlemmer et al. 2007].

Para [Messa 2010], as possibilidades disponíveis para utilização em AVAs são numerosas. O importante é dar-se conta de suas potencialidades em termos de aprimorar as atitudes pedagógicas

dos estudantes, dos professores e das instituições educacionais e, do uso apropriado desse recurso que pode proporcionar uma aprendizagem mais significativa para qualquer usuário, independente de idade desde que bem planejado por instituições educacionais.

2.2. Mineração de Dados Educacionais

A Mineração de Dados Educacionais, do inglês *Educational Data Mining* (EDM), é a aplicação de técnicas de *Data Mining* em dados provenientes de plataformas ou ambientes de educação on-line [Romero and Ventura 2007]. A EDM tem emergido como uma área de pesquisa nos últimos anos por pesquisadores em diversas áreas buscando analisar grandes conjuntos de dados, a fim de resolver as questões de investigação educacional [Baker and Yacef 2009].

Segundo [Baker and Yacef 2009], a EDM é a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar bases de dados coletadas em ambientes educacionais. Assim, é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o contexto na qual a aprendizagem ocorre, além de outros fatores que a influenciam.

O processo de EDM converte os dados brutos de sistemas educacionais em informação útil que pode ser usado por desenvolvedores, professores, pesquisadores educacionais, etc. Este processo não difere muito de outras áreas de aplicação de mineração de dados porque baseia-se nos mesmos passos do processo de mineração de dados em geral [Garcia et al. 2011]. São eles:

Passo 01 - Pré-processamento → Os dados obtidos com o ambiente educacional têm que primeiro ser pré-processados para transformá-los em um formato apropriado para a mineração. Algumas das principais tarefas do pré-processamento são: limpeza, seleção de atributos, os atributos de transformação, de integração de dados etc.;

Passo 02 - Mineração de dados (Extração de Padrões) → É o passo central que identifica todo o processo. Durante esta etapa, as técnicas de mineração de dados são aplicadas a dados previamente pré-processados.

Passo 03 - Pós-processamento → É a etapa final em que os resultados obtidos ou modelo são interpretados e usados para tomar decisões sobre o ambiente educacional.

Os dados gerados pelos alunos e instrutores em ambientes de *e-learning* podem fornecer rápidas e importantes compreensões acerca do desempenho, da motivação e do nível de participação dos alunos no curso. Essas compreensões podem sugerir mudanças no curso, intervenções significativas na metodologia ou mesmo um contato individual com alunos desmotivados ou com baixa interação [Romero et al. 2008].

2.3. Análise de Clusters

A análise de *clusters* é um procedimento com fundamentos da Estatística Multivariada que tenta agrupar um conjunto de dados em grupos homogêneos, chamados *clusters*, onde os dados podem ser objetos ou variáveis. Cada elemento pertencente a um determinado cluster é semelhante a todos os outros pertencentes a ele e é diferente dos elementos dos outros clusters [Faria et al. 2014].

Para [Soares et al. 2006], a clusterização é a técnica de agrupar dados (objetos) de uma base de dados de acordo com alguma medida de similaridade ou de dissimilaridade. Os métodos de clusterização podem ser classificados como hierárquicos e não hierárquicos, também conhecidos como particionais.

- **Métodos não Hierárquicos** - Dividem o conjunto de objetos de uma base de dados em vários subconjuntos disjuntos e procuram iterativamente o melhor particionamento até atingir uma condição de parada.

- **Métodos Hierárquicos** - O método hierárquico consiste em uma série de sucessivos agrupamentos ou sucessivas divisões de elementos, onde os elementos são agregados ou desagregados. Os resultados da clusterização hierárquica são representados através de uma árvore de clusters, também chamada de Dendograma. Neste trabalho serão apresentados os resultados para os agrupamentos utilizando um dos métodos hierárquicos.

Assim, o uso dos métodos de EDM em conjunto com softwares educacionais, torna possível apontar os diferentes fatores que influenciam o comportamento do aluno e identificar aspectos sutis, muitas vezes imperceptíveis do design de software, que instigam ou incentivam o surgimento de comportamentos indesejados e inadequados por parte dos alunos. Através desta verificação, a EDM também contribui para oferecer princípios de desenvolvimento que podem ser aplicados para criar softwares que minimizam o problema de comportamento e maximizam a aprendizagem do aluno [Baker et al. 2011].

2.4. Trabalhos relacionados

Com o objetivo de conhecer e entender diversos problemas relacionados aos Ambientes Virtuais e o processo de extração do conhecimento através mineração de dados nesses ambientes, foram listados alguns dos trabalhos na área que utilizaram a técnica de clusterização:

[Romero et al. 2008] compararam diferentes técnicas de mineração de dados e técnicas para classificar os alunos com base em seus dados de interação no moodle e na nota final obtida no curso. Eles desenvolveram uma ferramenta de mineração específica para facilitar a configuração e execução de técnicas de mineração de dados para os instrutores. Eles usaram dados reais de sete cursos do Moodle com estudantes da Universidade de Córdoba.

[Prass 2004] realizou um relevante estudo comparativo dos principais modelos de algoritmos de Análise de Agrupamento existentes na literatura e implementados em softwares, com o objetivo de usar no processo de descoberta de conhecimentos em grandes bancos de dados (*Knowledge Discovery in Databases - KDD*). O autor destacou que os algoritmos são diferenciados de acordo com o seu método de formação (Hierárquico, Partição, Baseado em Modelo, Baseado em Grade e Baseado em Densidade) e também pela medida de distância que expressa a similaridade ou dissimilaridade entre os objetos.

[Faria et al. 2014] propuseram um modelo de análise de dados, com base nos métodos cartas de controle, regressão logística e análise de clusters. O objetivo do trabalho foi extrair conhecimento relevante para a previsão do desempenho escolar, auxiliando os professores de ensino online a supervisionar de forma eficiente a atividade e o desempenho dos seus estudantes.

Com foco no desempenho, o trabalho de [França and Amaral 2013] apresentaram o uso de técnicas de clusterização para a mineração de dados, visando a formação de grupos similares de estudantes com dificuldades de aprendizagem no ensino de Programação. Com isso, esperava-se ser possível formular estratégias pedagógicas adequadas a grupos de aprendizes no intuito de melhorar o seu desempenho.

3. Realização do experimento

O objetivo do experimento foi realizar, a partir da utilização de mineração de dados educacionais, a descoberta de conhecimento relevante a partir da análise das interações e do desempenho dos alunos iniciantes na educação a distância em um ambiente virtual de aprendizagem.

3.1. Procedimentos

O experimento foi realizado na disciplina de Introdução à Educação a Distância (IEaD), no Bacharelado em Administração Pública por EAD, ofertado pela UNIVASF, para uma turma de 200 alunos

matriculados, distribuídos em 06 polos em cidades da região, onde foram coletados e analisadas as diversas informações contidas no banco de dados da disciplina dentro do ambiente Moodle, possibilitando a extração dos dados relevantes para a pesquisa.

A disciplina foi ofertada pelo ambiente Moodle e foi dividida em tópicos distribuídos ao longo de quatro semanas de duração. A estratégia de ensino do conteúdo foi baseada em vídeo-aulas, apostilas, fóruns de discussão, chats e troca de mensagens pelo ambiente. Além do professor, oito tutores online e seis tutores presenciais nos polos apoiaram o desenvolvimento das atividades da disciplina.

Ao longo de cada semana do curso, o material didático era disponibilizado, assim como um fórum específico para tratar do conteúdo visto foi criado para proporcionar interações entre professor, alunos e tutores. Foi realizado também um chat com os alunos por cada polo, onde foram esclarecidas as dúvidas sobre conteúdos da disciplina, a modalidade e o próprio curso.

Foram realizadas duas avaliações no curso, uma simulada, onde os alunos puderam ter o primeiro contato com uma avaliação online e a avaliação da disciplina, logo após a conclusão do conteúdo. Ao final da disciplina, 161 alunos foram aprovados, 15 reprovados por nota e 24 reprovados por não realizarem as atividades avaliativas.

3.2. Coleta e tratamento de dados

A coleta de dados e o posterior tratamento foi realizado utilizando a metodologia proposta por [Fayyad et al. 1996], dividida em três etapas principais: Preparação dos dados, extração de dados e o pós-processamento [Pimentel and Omar 2006].

A etapa de preparação dos dados é composta de três fases: Seleção, Pré-processamento e transformação dos dados. Na fase de seleção, os dados das interações dos alunos oriundos do Moodle foram coletados a partir de várias consultas SQL. Na fase de pré-processamento, é realizada a limpeza dos dados selecionados, a fim de garantir que sua qualidade e acurácia. Segundo [Olson and Delen 2008], a limpeza dos dados resume em filtrar ou atestar a presença de *outliers*, agregar e preencher os valores ausentes, (*missing values*), na tabela de dados. A transformação dos dados finaliza esta etapa, adequando ou normalizando os dados aos algoritmos selecionados, de acordo com a aplicação que será utilizada para a mineração de dados.

Na segunda etapa ocorre a extração dos padrões dos dados vindos da etapa anterior, através da técnica de mineração de dados e do algoritmo escolhido. Neste trabalho optou-se pela utilização da técnica de análise de *clusters*, utilizando a clusterização hierárquica, usando o método de Ward e a medida de distância euclidiana, com os dados sendo divididos em 4 grupos.

A terceira e última etapa é a de pós-processamento dos dados, que ocorre após a mineração de dados e é responsável pela avaliação, interpretação e utilização do conhecimento extraído, e caso se faça necessário, o retorno as etapas anteriores para as devidas adequações a fim de se obter uma mineração mais correta. Segundo [Lopes 2004], esta etapa consiste da fase de validação das descobertas efetuadas pela etapa de processamento dos dados e da visualização dos resultados encontrados.

4. Resultados e Discussões

Os resultados obtidos na construção deste experimento foram realizados no pacote estatístico R¹, a partir dos dados pré-processados nas etapas anteriores. Os atributos utilizados e suas respectivas descrições encontram-se na Tabela 1, assim como suas respectivas quantidades mínimas, máximas e médias encontradas.

¹Disponível em <http://www.r-project.org>

Tabela 1. Descrição dos atributos utilizados da Base de Dados do experimento

| Atributo | Descrição | Mínimo | Máximo | Média |
|----------------------|---|--------|--------|-------|
| Nota | Nota na avaliação final do curso | 0 | 100 | 71,26 |
| Chat | Quantidade de postagens em Chats | 0 | 46 | 6,61 |
| Download | Quantidade de Downloads realizados | 0 | 24 | 4,50 |
| Forum | Quantidade de postagens em fóruns | 0 | 50 | 4,06 |
| Msg_rec_do_professor | Quantidade de mensagens recebidas do professor | 0 | 13 | 2,11 |
| Msg_env_ao_professor | Quantidade de mensagens enviadas para o professor | 0 | 12 | 0,67 |
| Msg_rec_de_alunos | Quantidade de mensagens recebidas de outros alunos | 0 | 20 | 1,57 |
| Msg_env_para_alunos | Quantidade de mensagens enviadas para outros alunos | 0 | 51 | 1,62 |
| msg_rec_dos_tutores | Quantidade de mensagens recebidas dos tutores | 0 | 70 | 29,39 |
| Msg_env_aos_tutores | Quantidade de mensagens enviadas para os tutores | 0 | 97 | 13,67 |

Percebe-se que média da nota da turma ficou acima da média para aprovação, o que é reforçada pelo índice de aprovação da turma, aproximadamente 80% dos alunos. Merece destaque também a quantidade média de mensagens trocadas entre alunos e tutores, indicando um boa interação entre esses atores, bem acima da média das interações professor-aluno.

4.1. Matriz de correlação entre as variáveis

Após as etapas de preparação dos dados, foi realizada uma análise estatística antes de proceder com a mineração dos dados, com o objetivo de se buscar uma possível correlação da nota com as demais variáveis de interação. Foi obtida a matriz de correlação entre as variáveis coletadas e os resultados obtidos estão exibidos na Tabela 2.

Tabela 2. Matriz geral de correlação entre as variáveis

| Matriz de Correlação: Pearson | Nota | Chat | Download | Forum | Msg_rec_do _professor | Msg_env_ao _professor | Msg_rec de_alunos | Msg_env_ para_alunos | msg_rec_ dos_tutores | Msg_env_ aos_tutores |
|----------------------------------|--------|--------------|----------|--------------|--------------------------|--------------------------|----------------------|-------------------------|-------------------------|-------------------------|
| Nota | 1,000 | | | | | | | | | |
| Chat | 0,203 | 1,000 | | | | | | | | |
| Download | 0,041 | 0,174 | 1,000 | | | | | | | |
| forum | 0,188 | 0,439 | 0,193 | 1,000 | | | | | | |
| Msg_rec_do_professor | -0,067 | 0,017 | 0,136 | 0,065 | 1,000 | | | | | |
| Msg_env_ao_professor | -0,009 | 0,094 | 0,099 | 0,131 | 0,597 | 1,000 | | | | |
| Msg_rec_de_alunos | 0,059 | 0,105 | -0,070 | 0,162 | 0,044 | 0,140 | 1,000 | | | |
| Msg_env_para_alunos | 0,011 | 0,015 | -0,071 | 0,097 | 0,175 | 0,235 | 0,777 | 1,000 | | |
| Msg_rec_dos_tutores | 0,051 | 0,115 | 0,125 | 0,168 | 0,044 | 0,090 | 0,203 | 0,152 | 1,000 | |
| Msg_env_aos_tutores | 0,084 | 0,249 | 0,169 | 0,583 | 0,142 | 0,371 | 0,165 | 0,177 | 0,217 | 1,000 |

Podemos observar na matriz que a correlação entre os atributos de interação e o atributo notas é fraca, implicando que a quantidade de interações nesta turma não foi um fator que influenciou na nota dos estudantes. Foram destacados na tabela alguns campos onde houve alguma correlação entre os dados.

É possível citar uma correlação moderada entre os atributos chat e fórum, indicando que a quantidade de interações no fórum está relacionada a participação dos alunos no chat, e entre postagens no fórum e mensagens enviadas aos tutores. Ainda uma correlação moderada entre as mensagens enviadas e recebidas do professor e uma correlação forte entre as mensagens recebidas e enviadas entre os alunos, evidenciando que a interação nesses casos, foi nos dois sentidos: o receptor respondia com frequência ao emissor da mensagem, o que é um aspecto positivo da interação, pois indica a presença de diálogos.

Como a análise da correlação não indicou possíveis influências das demais variáveis na nota final do aluno, não foi recomendável, para esses dados, utilizar técnicas de predição de desempenho, como por exemplo um modelo a partir de uma regressão linear. Assim, optou-se pela utilização das técnicas de clusterização na mineração e análise dos dados, a fim de se descobrir grupos de comportamentos diferenciados em relação às interações dos alunos dessa turma.

4.2. Análise de agrupamentos

Foram realizadas quatro clusterizações, duas hierárquicas e duas não-hierárquicas, para formação dos grupos conforme as características dos dados coletados. Este trabalho apresenta uma clusterização hierárquica para quatro grupos. As demais são encontradas em [Silva 2015].

A partir dos dados normalizados, um total de 10 atributos e 169 registros. A clusterização foi realizada utilizando como parâmetros a distância euclidiana e o método de Ward. A opção por esse método e essa medida de distância foi em razão das mesmas apresentarem os melhores resultados para os dados coletados. Os grupos obtidos tiveram os seguintes tamanhos:

- **Grupo 1** → 9 instâncias (Alunos);
- **Grupo 2** → 96 instâncias;
- **Grupo 3** → 26 instâncias;
- **Grupo 4** → 38 instâncias.

Em seguida, foi realizada a sumarização dos dados de cada grupo. A tabela 3 apresenta a relação entre a média e o desvio padrão para cada grupo de acordo com cada atributo coletado.

Tabela 3. Tabela de Sumarização de Dados - Clusterização Hierárquica, K=4

| Variáveis | Grupo 01 | | Grupo 02 | | Grupo 03 | | Grupo 04 | |
|----------------------|----------|-------------|----------|-------------|----------|-------------|----------|-------------|
| | Média | Desv Padrão | Média | Desv Padrão | Média | Desv Padrão | Média | Desv Padrão |
| Chat | 6,67 | 6,34 | 4,43 | 5,41 | 22,69 | 12,30 | 1,08 | 2,02 |
| Download | 2,56 | 2,24 | 4,89 | 4,14 | 6,50 | 4,95 | 2,60 | 2,27 |
| Forum | 4,10 | 1,83 | 3,74 | 3,11 | 9,04 | 9,21 | 1,45 | 1,95 |
| Msg_rec_do_professor | 2,33 | 1,73 | 1,69 | 1,32 | 3,23 | 2,64 | 2,36 | 1,03 |
| Msg_env_ao_professor | 1,00 | 1,73 | 0,32 | 0,75 | 2,62 | 3,44 | 0,16 | 0,37 |
| Msg_rec_de_alunos | 11,67 | 4,33 | 0,91 | 1,57 | 2,35 | 2,30 | 0,34 | 0,75 |
| Msg_env_para_alunos | 15,78 | 13,67 | 0,67 | 1,50 | 2,35 | 2,93 | 0,16 | 0,50 |
| Msg_rec_dos_tutores | 42,89 | 20,46 | 27,22 | 10,80 | 37,31 | 17,21 | 26,26 | 5,94 |
| Msg_env_aos_tutores | 13,56 | 7,16 | 11,99 | 8,50 | 27,62 | 20,91 | 8,42 | 6,30 |
| Nota | 71,81 | 26,10 | 77,70 | 16,78 | 78,75 | 14,08 | 49,74 | 15,78 |

Podemos observar nos dados na tabela algumas características peculiares para cada grupo. No grupo 01, além de uma média satisfatória das notas, destaca-se a quantidade de mensagens recebidas dos tutores e a quantidade de mensagens recebidas e enviadas para outros alunos, sendo o grupo com maior quantidade em ambos os atributos, o que representa os alunos com boa interação entre si e com os tutores.

Pertencem ao grupo 02 os alunos com segunda maior média entre os grupos, próxima da maior média encontrada (Grupo 03). Neste grupo, nenhuma das interações dos alunos se destaca em comparação com as dos outros grupos, tendo apenas uma média satisfatória de interações nas mensagens recebidas e enviadas aos tutores. Mais uma vez, a interação tutor-aluno pode ter contribuído para o desempenho dos alunos.

No grupo 03 são encontrados os alunos com a maior média de notas. Este grupo é o que possui a maior média das interações no Chat, Fórum e nas interações através de mensagens com o professor, de mensagens enviadas aos tutores, maior quantidade de downloads e o segundo maior nas mensagens recebidas dos tutores. As interações diretas (mensagens) entre os alunos desse grupo foram as menores em relação aos demais grupos, mas que foi devidamente compensada pelo nível de interação nos fóruns e nos chats desses estudantes.

Já o grupo 04 é o que possui a menor média no atributo nota. Os alunos deste grupo são os que tiveram as menores médias nas interações em relação aos demais grupos, o que poderia justificar o baixo rendimento no quesito desempenho na avaliação.

Em todos os grupos, as interações entre alunos e tutores sobressaíram em relação às demais, indicando que a participação dos tutores no processo de ensino e aprendizagem foi bem ativa.

As análises desse agrupamento são úteis para os coordenadores de curso e tutoria, além de professores e os próprios alunos, na medida em que apontam indicadores importantes para busca

de experiências válidas (como a interação entre alunos e tutores) de modo a replicá-las em outras turmas. Também serve como referência para identificação de possíveis problemas por parte dos alunos (os dos grupo 04), onde tanto a nota quanto as interações foram insatisfatórias.

4.3. Análise da correlação das variáveis nos grupos formados

A partir dos grupos encontrados, foram realizados testes de correlação para determinar a relação entre os atributos de cada grupo. Para as células da matriz de correlação entre os atributos que possuem uma correlação positiva, implica que o crescimento do valor de atributo está ligado diretamente ao crescimento do outro correlacionado a ele. As células que possuem uma relação negativa, implica que o crescimento de um atributo esta inversamente ligado ao crescimento do outro a ele correlacionado.

Tabela 4. Matriz de correlação - Grupo 01

| Matriz de Correlação: Pearson | Chat | Download | Forum | Msg_rec_do_ professor | Msg_env_ao_ professor | Msg_rec_ de_alunos | Msg_env_ para_alunos | msg_rec_ dos_tutores | Msg_env_ aos_tutores |
|----------------------------------|---------------|---------------|--------------|--------------------------|--------------------------|-----------------------|-------------------------|-------------------------|-------------------------|
| Chat | 1,000 | | | | | | | | |
| Download | -0,240 | 1,000 | | | | | | | |
| forum | -0,029 | 0,470 | 1,000 | | | | | | |
| Msg_rec_do_professor | -0,159 | -0,118 | 0,420 | 1,000 | | | | | |
| Msg_env_ao_professor | 0,478 | -0,290 | 0,472 | 0,625 | 1,000 | | | | |
| Msg_rec_de_alunos | -0,382 | -0,558 | -0,247 | -0,067 | -0,233 | 1,000 | | | |
| Msg_env_para_alunos | -0,237 | -0,526 | -0,084 | 0,737 | 0,238 | 0,427 | 1,000 | | |
| Msg_rec_dos_tutores | -0,404 | 0,383 | -0,290 | -0,383 | -0,469 | -0,065 | -0,362 | 1,000 | |
| Msg_env_aos_tutores | -0,177 | -0,208 | 0,471 | 0,376 | 0,343 | 0,648 | 0,420 | -0,255 | 1,000 |
| Nota | -0,039 | 0,194 | 0,387 | 0,158 | 0,038 | -0,294 | -0,148 | -0,276 | -0,238 |

Na matriz do grupo 01 (Tabela 4), foi possível observar correlações positivas moderadas e fortes entre os atributos, indicando que os alunos desse grupo interagiram de maneira satisfatória com o professor, o conteúdo e entre si. Vale ressaltar que esse foi o grupo de menor tamanho (09 instâncias), o que pode ter influenciado nessas características relevantes do grupo.

Para os grupos 02 e 04, as suas matrizes de correlação apresentaram somente uma correlação moderada e forte, respectivamente, entre os atributos mensagens enviadas e recebidas entre os alunos (*Grupo 03* = 0,778 e *Grupo 04* = 0,656). A correlação entre os outros atributos se mostrou fraca, demonstrando pouca relação entre esses atributos em ambos os grupos. Foram os grupos com maiores quantidades de elementos (96 e 38 respectivamente), mas a única característica que sobressaiu entre os seus integrantes foi a troca de mensagens entre alunos.

Tabela 5. Matriz de correlação - Grupo 03

| Matriz de Correlação: Pearson | Chat | Download | Forum | Msg_rec_do_ professor | Msg_env_ao_ professor | Msg_rec_ de_alunos | Msg_env_ para_alunos | msg_rec_ dos_tutores | Msg_env_ aos_tutores |
|----------------------------------|---------------|----------|--------------|--------------------------|--------------------------|-----------------------|-------------------------|-------------------------|-------------------------|
| Chat | 1,000 | | | | | | | | |
| Download | -0,342 | 1,000 | | | | | | | |
| forum | 0,090 | 0,111 | 1,000 | | | | | | |
| Msg_rec_do_professor | -0,546 | 0,306 | -0,160 | 1,000 | | | | | |
| Msg_env_ao_professor | -0,685 | 0,000 | -0,205 | 0,753 | 1,000 | | | | |
| Msg_rec_de_alunos | -0,088 | 0,245 | 0,170 | 0,131 | 0,048 | 1,000 | | | |
| Msg_env_para_alunos | -0,505 | 0,264 | 0,074 | 0,222 | 0,471 | 0,660 | 1,000 | | |
| Msg_rec_dos_tutores | -0,397 | 0,361 | -0,021 | 0,049 | -0,070 | -0,153 | 0,114 | 1,000 | |
| Msg_env_aos_tutores | -0,422 | 0,009 | 0,633 | 0,070 | 0,236 | -0,090 | 0,275 | 0,275 | 1,000 |
| Nota | 0,202 | 0,174 | -0,016 | -0,305 | -0,329 | 0,099 | 0,097 | 0,160 | -0,033 |

No grupo 03 (Tabela 5), destaca-se na matriz uma correlação positiva forte entre as mensagens trocadas entre professor e alunos e uma correlação moderada entre os atributos: postagens no fórum com mensagens enviadas aos tutores, mensagens enviadas para alunos com mensagens enviadas ao professor e com mensagens recebidas de alunos.

Interessante observar as correlações negativas moderadas entre as postagens do chat e as mensagens enviadas ao professor, tutor e outros alunos. Isso evidencia uma relação inversa entre as postagens no chat e o envio de mensagens para o professor, tutor e colegas. Como esse grupo foi o que em média, mas postou em chats, isso pode indicar que o alunos desse grupo utilizaram-se mais desse recurso do que o de enviar mensagens.

Por fim, a correlação entre as notas e os outros atributos em todas as matrizes de correlação para os quatro clusters se mostrou fraca ou muito fraca, demonstrando pouca correlação entre os atributos e as notas dos alunos.

Cada grupo pode conter características importantes para o planejamento e a execução do curso. Embora sejam alunos iniciantes e os grupos formados podem sofrer mudanças ao longo do curso, a análise atual de cada grupo pode indicar a necessidade de algum ajuste no processo de ensino-aprendizagem ou mesmo no design instrucional do curso. Medidas podem ser tomadas para aumentar a colaboração entre os alunos, facilitar o acesso ao conteúdo, aumentar a participação nos chats e fóruns, baseadas nos resultados das análises dos grupos gerados.

5. Considerações finais e trabalhos futuros

Este estudo teve como objetivo principal a descoberta de conhecimento a respeito do desempenho dos alunos através da análise de suas interações com o uso da mineração dos dados oriundos da base de dados do Moodle-SEAD/UNIVASF.

A técnica escolhida para esta pesquisa foi a clusterização hierárquica para a análise dos dados, onde foi possível observar que em alguns grupos com maiores médias de notas, uma quantidade de interações relativamente alta e em alguns grupos menores médias de notas, valores de interação baixos, o que nos permite afirmar que para esses grupos a quantidade de interações foi um fator que influenciou na nota final do aluno. A técnica se mostrou apropriada para se ter uma visão geral dos alunos divididos em grupos por similaridade entre seus elementos.

Os resultados da clusterização apontam grupos com características distintas entre si, mas ainda com alguma similaridade, como por exemplo, as interações aluno-tutor. De uma maneira geral, cada grupo formado pode oferecer alguma particularidade para que a coordenação do curso ou o professor possa promover ajustes no curso, como por exemplo, promovendo tarefas em equipe que estimulem o trabalho colaborativo e por consequência, eleve ainda mais o nível de interação entre os alunos.

Problemas de baixo desempenho, como os alunos do Grupo 04, podem ser melhor estudados suas causas e relações e, a partir daí, medidas podem ser tomadas no sentido de melhorar esse desempenho.

Planeja-se em trabalhos futuros resolver alguns desafios que surgem a partir dessa construção, com a aplicação das técnicas de clusterização na mesma turma, a fim de averiguar se os atributos estudados nesse trabalho passaram a influenciar no desempenho e nas interações. A aplicação do processo em outras turmas e o uso de outras tarefas de EDM nessa mesma turma podem ser feitos a fim de descobrir novos comportamentos ou prever comportamentos futuros dos alunos. Podem ser agregadas técnicas de visualização gráfica para geração de gráficos interativos e de fácil de compreensão e assim como a construção de uma ferramenta que simplifique e torna mais transparente a utilização desse processo por professores e gestores dos cursos.

6. Agradecimentos

Essa pesquisa é apoiada pela FACEPE-APQ No. 0525-1.03/14. Os também autores agradecem o apoio da PRPPGI/UNIVASF (Edital 18/2013) e à SEAD/UNIVASF por colaborar e oferecer o ambiente para a pesquisa. Alex Sandro Gomes é bolsista DT Nível 2/CNPq, processos nº 310466/2012-1 e nº 475634/2013-6.

Referências

Anderson, T. (2008). The theory and practice of online learning. Athabasca University Press.

- Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, 19(02).
- Baker, R. S. J. D. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*.
- Clark, R. C. and Mayer, R. E. (2011). *E-learning and the science of instruction*. John Wiley & Sons.
- Faria, S. M. S. M. L. et al. (2014). *Educational data mining e learning analytics na melhoria do ensino online*.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*.
- França, R. S. d. and Amaral, H. J. C. d. (2013). Mineração de dados na identificação de grupos de estudantes com dificuldades de aprendizagem no ensino de programação. *RENOTE*, 11(1).
- Garcia, E., Romero, C., Ventura, S., and de Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education*, 14(2).
- Gonçalves, V. (2007). *e-Learning: reflexões sobre cenários de aplicação*. Universidade da Madeira, Sociedade Portuguesa de Ciências da Educação.
- Lopes, M. C. S. (2004). *Mineração de dados textuais utilizando técnicas de clustering para o idioma português*. PhD thesis, Universidade Federal do Rio de Janeiro.
- Messa, W. C. (2010). Utilização de ambientes virtuais de aprendizagem-avas: A busca por uma aprendizagem significativa. *Revista Brasileira de Aprendizagem Aberta e a Distância*, 9.
- Olson, D. L. and Delen, D. (2008). *Advanced data mining techniques*. Springer.
- Pimentel, E. P. and Omar, N. (2006). Descobrendo conhecimentos em dados de avaliação da aprendizagem com técnicas de mineração de dados. *Anais do Workshop de Informática na Escola*, 1(1).
- Prass, F. S. (2004). Estudo comparativo entre algoritmos de análise de agrupamentos em data mining. 2004. 71 f. Master's thesis, Universidade Federal de Santa Catarina, Florianópolis, SC.
- Rodrigues, R. L., Ramos, J. L. C., Silva, J. C. S., and Gomes, A. S. (2014). A literatura brasileira sobre mineração de dados educacionais. In *Anais do Workshop de Mineração de Dados em Ambientes Virtuais do Ensino/Aprendizagem - (CBIE 2014, pg 621-630)*.
- Romero, C. and Ventura, S. (2007). *Educational data mining. Expert systems with applications*, 33(1).
- Romero, C., Ventura, S., Espejo, P. G., and Hervás, C. (2008). Data mining algorithms to classify students. In *EDM*.
- Schlemmer, E., Saccol, A. Z., and Garrido, S. (2007). Um modelo sistêmico de avaliação de softwares para educação a distância como apoio 'a gestão de ead. *REGE Revista de Gestão*, 14(1).
- Silva, R. E. D. (2015). Análise das interações de estudantes na educação a distância a partir de uma técnica de mineração de dados educacional. *Trabalho de Conclusão de Curso.(Graduação em Engenharia da Computação) - Universidade Federal do Vale do São Francisco - UNIVASF*.
- Soares, S., Ochi, L. S., Drummond, L., et al. (2006). Um algoritmo de construção e busca local para o problema de clusterização de bases de dados. *TEMA-Tendências de Matemática Aplicada e Computacional*, 7(1):109–118.