

Mineração de Dados na Descoberta do Padrão de Usuários de um Sistema de Educação à Distância

Janaina R Penedo, Eliane P Capra

Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur, 296 – Urca - CEP 22290-240 – Rio de Janeiro – RJ – Brazil
{janaina.penedo, eliane.capra}@uniriotec.br

Abstract. *As other systems of information, the systems used in the E-Learning generate great amount of data little explored. In this perspective, it appears the necessity of transforming these data into useful information. This article describes a process of discovery of useful knowledge in a system of E-Learning with intention to investigate the excellent information that assist in the identification of the standard of the users who use the system.*

Resumo. *Assim como outros sistemas de informação, os sistemas utilizados na educação à distância (EAD) geram grande quantidade de dados pouco explorados. Nessa perspectiva surge a necessidade de se transformar estes dados em informações úteis. Este artigo descreve um processo de descoberta de conhecimento útil em um sistema de EAD com intuito de investigar as informações relevantes que auxiliem na identificação do padrão dos usuários que utilizam o sistema.*

1. Introdução

A cada ano mais operações estão sendo computarizadas, todas acumulando dados em operações, em atividades e em desempenho. Todos estes dados “prendem” valiosas informações, como tendências e padrões, que podem ser usados para melhorar ou aperfeiçoar decisões de negócio [Goebel 1999].

As ferramentas e técnicas empregadas para análise automática e inteligente destes de dados são os objetos tratados pelo campo emergente da Descoberta de Conhecimento em Bancos de Dados (DCBD), da expressão em inglês *Knowledge Discovery in Databases* (KDD) [Silva 2011]. No presente artigo será utilizada a expressão KDD.

Os objetivos da descoberta do conhecimento são definidos pelo que se deseja de retorno do sistema. Podem-se distinguir dois tipos de objetivos: Verificação e Descoberta. Na Verificação o objetivo se limita a verificar as hipóteses. Na descoberta o objetivo é automaticamente encontrar padrões. A Descoberta ainda pode ser dividida em Predição, onde o sistema encontra padrões para prever o futuro comportamento de algumas entidades, e Descrição, onde o sistema encontra padrões de forma que facilitem o entendimento para o usuário. Em resumo, a Predição envolve o uso de algumas variáveis ou campos do banco de dados para prever valores desconhecidos ou futuras variáveis de interesse. A Descrição foca em encontrar padrões que possam ser interpretados pelos usuários [Fayyad 1996].

As técnicas de KDD são utilizadas para, automaticamente, identificar informações úteis em grandes massas de dados. Encontrar padrões nessas massas envolve: mineração de dados (*data mining*), extração do conhecimento, descoberta da informação e padrão de processamento de dados [Fayyad 1996].

Além disso, os resultados obtidos com a utilização das técnicas de KDD podem ser utilizados no gerenciamento de informações, processamento de pedidos de informação, tomada de decisão, controles de processos e em muitas outras aplicações [Dias 2002].

Muitas são as áreas que utilizam as técnicas de KDD como facilitador para reconhecimento de padrões ou modelos em suas bases de informações. A área de Ensino, especificamente a Educação à Distância, é uma candidata natural para esse processo.

O presente artigo, de caráter quantitativo e exploratório, teve como objetivo aplicar as técnicas de KDD em uma base de dados contendo informações dos acessos de alunos em um sistema de Educação à Distância (EAD) para identificar o padrão dos usuários que melhor se adaptam ao sistema de EAD disponibilizado.

2. Trabalhos Relacionados

Alguns trabalhos encontrados na literatura dizem respeito à utilização de processos de mineração de dados na área da educação. Tais trabalhos estão descritos a seguir:

- a) Santos (2002) utilizou técnicas de mineração de dados para descobrir e analisar as interações de usuários com um ambiente de ensino à distância analisando a navegação dos usuários. Com o resultado do trabalho o autor contribuiu com uma taxonomia de acessos para associar padrões na execução de atividades propostas durante o curso e análise de melhores *templates* de navegação.
- b) Garcia (2007) efetuou a mineração de dados com objetivo de aperfeiçoar conteúdos de um portal educacional por meio da descoberta dos conteúdos que mais interessam aos usuários. Isso proporcionou *feedback* aos autores dos cursos.
- c) Romero *et al.* (2008) desenvolveram uma ferramenta para minerar e apresentar visualmente dados de navegação dos usuários. A mineração de dados foi utilizada para descobrir padrões de sequências de trilhas visitadas. Como resultado, os autores apresentaram um grafo em que as páginas *web* são os nodos e as linhas indicam o sentido do acesso ou navegação entre um nodo e outro, com números indicando a probabilidade de cada ligação. O objetivo é o auxílio aos professores e desenvolvedores de materiais educacionais na interpretação dos dados.
- d) Oliveira (2005) utilizou a mineração de dados em um estudo sobre o perfil dos candidatos que se inscreveram em um processo seletivo. Os resultados permitiram identificar diferentes públicos, relacionando regiões de origem, poder aquisitivo, meios de comunicação mais eficazes e motivos mais significativos. O objetivo era o

auxílio a melhores ações de marketing voltadas a divulgação da instituição, especificamente do processo seletivo.

- e) Amorim (2008) utilizou técnicas de mineração de dados para prever a evasão em cursos à distância. Foram consideradas informações acadêmicas gerais e financeiras dos alunos. O objetivo do autor de Amorim tinha cunho administrativo, com foco em descobrir padrões de risco de abandono nos cursos de EAD.

3. Método de Pesquisa

A pesquisa de caráter quantitativo e exploratório foi realizada com base em um estudo de caso. O uso de estudos de caso pode ser definido ao tentar conhecer com profundidade “como” e “por que” tal fenômeno ocorre [Stake 1995]. No caso da presente pesquisa, pretende-se responder questões do tipo “como?” o processo de KDD pode auxiliar na identificação de padrões ou modelos de comportamento dos alunos do sistema de EAD.

Para tal, optou-se por um estudo de caso único. Para realização do estudo de caso, foram utilizadas as etapas propostas por Fayyad [Fayyad 1996], descritas a seguir:

Etapa 1: Entender o domínio aonde será aplicado o processo de KDD, identificando os objetivos do ponto de vista do cliente.

Etapa 2: Selecionar uma série de dados aonde será executado o processo.

Etapa 3: Limpeza e Pré-processamento dos dados.

Etapa 4: Transformação dos dados.

Etapa 5: Objetivos versus o método de mineração de dados a ser utilizado.

Etapa 6: Escolha do algoritmo de mineração de dados e o método de seleção a ser utilizado na busca pelos padrões.

Etapa 7: Mineração de dados em si.

Etapa 8: Interpretação dos padrões.

Etapa 9: Utilização das informações obtidas no processo, incorporando o conhecimento em outro sistema para ações futuras ou simplesmente documentando e reportando as partes interessantes.

4. Resultados do Estudo de caso

O processo de KDD foi aplicado em uma amostra de dados reais referentes ao *log* de acessos do sistema de EAD utilizado na Fundação Centro de Ciências e Educação Superior à Distância do Estado do Rio de Janeiro - Fundação CECIERJ/Consórcio CEDERJ [Cederj 2011]. Esse Consórcio reúne seis universidades públicas sediadas no Estado do Rio de Janeiro [Cederj 2011]. O banco de dados disponibilizado reuniu informações referentes aos acessos efetuados ao sistema durante todo o ano de 2010.

O aluno do Consórcio CEDERJ está matriculado em uma das universidades desse consórcio, dependendo do curso e do polo regional a que esteja vinculado. Os polos são os locais onde os alunos matriculados podem realizar as atividades presenciais

obrigatórias, como aulas de laboratório, avaliações, entre outros. Atualmente, o Consórcio possui mais de 30 mil alunos matriculados em diversos cursos de graduação (Computação, Matemática, Ciências Biológicas, Física, Administração, Química, História, Pedagogia, Turismo e Licenciatura em Pedagogia).

Nas seções seguintes são descritos os resultados obtidos em cada etapa baseada na proposta de Fayyad [Fayyad 1996].

4.1. Etapa 1: Entender o domínio

Por se tratar de uma pesquisa acadêmica, considerou-se como principal domínio do problema a identificação do padrão de usuários que acessam um sistema de Educação a Distância.

4.2. Etapa 2: Extração dos Dados

Como a extração dos dados no processo de KDD pode visar aplicação [Dilly 1995], nesta etapa foi necessário envolver a integração e compatibilização de bases de dados.

Dessa forma, além dos dados extraídos da base de dados SQL (*log* contendo os acessos dos usuários no ano de 2010), também foram obtidos outros dados para análise. Foram extraídas as seguintes informações: dados cadastrais dos usuários (matrícula, sexo, idade, CR, situação, município de origem, curso, universidade, polo), tipos de acessos dentro do sistema (data, hora, disciplina, ferramenta utilizada, ação, perfil, matrícula, curso, polo) e dados de carga gerada no sistema (*in*, *out*, data, hora).

Para que os dados dos *log* fossem unificados, foi necessário realizar a junção dessas informações. Essa junção foi executada utilizando o campo matrícula do usuário, obtido dos dados cadastrais.

4.3. Etapa 3: Limpeza e Pré-processamento

Nesta etapa foi verificada a parte de inconsistência dos dados no nível de valores ausentes, discrepância de valores, entre outros. Levando-se em conta o tamanho da base de dados a ser analisada (acima de 2 milhões de linhas) houve a necessidade de utilizar um aplicativo para auxiliar na limpeza dos dados.

O aplicativo selecionado para realização desta pesquisa foi o *Waikato Environment for Knowledge Analysis* (WEKA), que é implementado em Java. Este aplicativo é gratuito e possui documentação completa de todos os algoritmos de mineração de dados implementados. A ferramenta possui funcionalidades desde o pré-processamento até a etapa de mineração dos dados, contemplando diversos algoritmos e métodos [Witten 1999].

A primeira etapa na utilização da ferramenta é a preparação do arquivo contendo os dados selecionados. A ferramenta WEKA trabalha com arquivos no formato *Attribute-Relation File Format* (ARFF), formato padrão de arquivo texto utilizado para representar dados [Oliveira 2005].

A primeira limpeza necessária encontrada na geração do arquivo foi a retirada das linhas onde o sistema não gerou o atributo “cod_ferramenta”. É através desse atributo que identificamos as ações dos usuários quando na utilização do sistema. Também foi necessário efetuar uma limpeza no campo “DataNascimento” e “sexo”, onde foram encontrados dados sem preenchimento (0,005% do total de dados).

4.4. Etapa 4: Transformação dos Dados

Nesta etapa é executada a padronização dos dados para suprir possíveis limitações nas técnicas da etapa de mineração de dados. Optou-se por aplicar, inicialmente, a transformação em uma amostra menor de dados. Foram utilizados os dados contidos no mês de Março/2010 (mês com maior número de acessos a plataforma). De acordo com o ambiente estatístico SStatNet [Nassar *et al.* 2011] a quantidade selecionada de dados é considerada adequada para representar a população de dados da pesquisa, sendo superior a necessária para um nível de confiança de 99,9% (271), conforme cálculo de tamanho da amostra (Figura 1).

Estimação de Percentual	
Tamanho da População	1015179
Precisão da Estimativa	50% ± 10%%
Nível de Confiança	95%
Tamanho da Amostra	96
Perda Amostral	Nenhuma

Para outros Níveis de Confiança	
Nível de Confiança	Tamanho da Amostra
99.9%	271
99%	166
90%	68

Figura 1 – Cálculo do tamanho da amostra SStatNet

Nesta etapa efetuou-se então a modificação de alguns campos na intenção de discretizar (criar faixas de dados) a base de dados.

Para o campo CR (Coeficiente de Rendimento), criou-se as seguintes faixas: RUIIM – CR entre 0.0 e 4,99; REGULAR – CR entre 5.0 e 6,99; BOM – CR acima de 7.0.

Para o campo Idade criou-se a faixa: IDADE1 – Idade entre 0 e 28 anos; IDADE2 – Idade acima de 28 anos.

Para o campo Ferramentas, foram criadas faixas de acordo com o tipo de ferramenta disponibilizada na plataforma. A plataforma CEDERJ disponibiliza a seus usuários 81 diferentes ferramentas a serem utilizadas. Destas, somente 27 são realmente utilizadas pelos usuários, conforme apresentado na Tabela 1:

Tabela 1. Faixas criadas para o campo “ferramentas”

Faixa	Tipo de Ferramentas
Administrativas	Login, Logout, Perfil, Sair SI. Disciplina
Disciplinas	SI. Disciplina, Material Complementar, Gerenciador de Arquivos, Participantes, Aulas na Web, Guia de Disciplina, Calendário das Práticas, Atividades, Avaliações e Gabaritos, Exercícios Complementares, Mat. Didático
Comunicação	SI Tutoria, SI conferência, Aviso, Mural, Mural dos polos, Mural do Curso, Fale conosco
Aplicativos	Forum, E-mail, Blog, Wiki

Foi necessário discretizar também campos relacionados a data, hora e quantidade de acessos. Para o campo data, dividiu-se o mês em duas quinzenas: quinzena1 (até o dia 15, inclusive) e quinzena2 (do dia 16 em diante). Para o campo hora, dividiu-se em 4 faixas: madrugada (de 23h às 05:59h), manhã (de 6h às 11:59h), tarde (de 12h às 17:59h) e noite (de 18h às 22:59h). O campo acesso diz respeito à quantidade de requisições enviadas a plataforma durante o mês. Neste caso, foram criadas as faixas: pequeno (até 20 requisições/mês), médio (de 20 à 50 requisições/mês), grande (de 50 a 100 requisições/mês) e superior (acima de 100 requisições / mês).

4.5. Etapa 5: Objetivos X Mineração de Dados

Esta etapa tratou da combinação dos objetivos versus o método de mineração de dados a ser utilizado.

De acordo com o objetivo proposto, a tarefa de mineração de dados foi enfocada pela descoberta de regras de associação. Regras de associação podem ser utilizadas para encontrar relacionamentos ou padrões frequentes em um conjunto de dados [Silva *et al.* 2003]. Com a associação pretende-se identificar relações entre os atributos ou a presença de padrões entre os dados que possam influenciar a tomada de decisões.

4.6. Etapas 6 e 7: Escolha do Algoritmo e Mineração de Dados

Optou-se por unir essas duas etapas tendo em vista que a escolha do algoritmo, no caso desta pesquisa, está diretamente vinculada à mineração dos dados.

O algoritmo escolhido para minerar os dados do sistema foi o denominado *Apriori* [Agrawal *et al.* 1994]. O algoritmo foi escolhido por ser citado como o de melhor desempenho dentre os algoritmos de associação [Wu *et al.*, 2007]. Esse algoritmo efetua diversos tipos de busca no banco de dados à procura de associações entre itens. É importante ressaltar que o algoritmo selecionado aceita campos nominais e

não aceita campos com atributos numéricos [Benicasa *et al.* 2006]. Ao executar as varreduras na base de dados, o algoritmo gera um conjunto de itens “candidatos” e, em seguida, verifica se os “candidatos” satisfazem o suporte mínimo estipulado [Silva *et al.* 2003].

4.7. Etapa 8: Interpretação dos Padrões

Esta etapa envolve a visualização dos dados extraídos do modelo a fim de obter resultados aos objetivos propostos. A Figura 2 apresenta o número de usuários ativos na plataforma de acordo com os cursos oferecidos no mês de março/2010.

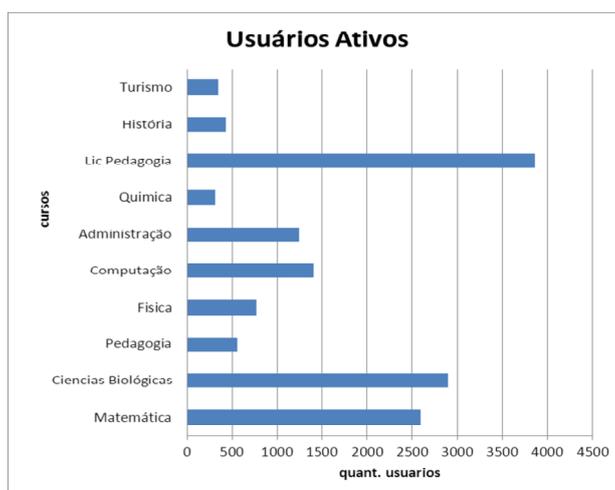


Figura 2 – Usuários Ativos na Plataforma EAD – Mês Março/2010

Utilizando para correlação o atributo “Sexo”, pode-se encontrar diferenças entre os usuários dos cursos. Os cursos referentes à Pedagogia, Licenciatura em Pedagogia possuem predominância de alunos do sexo feminino. O curso de Turismo é o que apresenta maior diversidade de utilização de ambos os sexos com uma proporção de 60/40 (Figura 3).

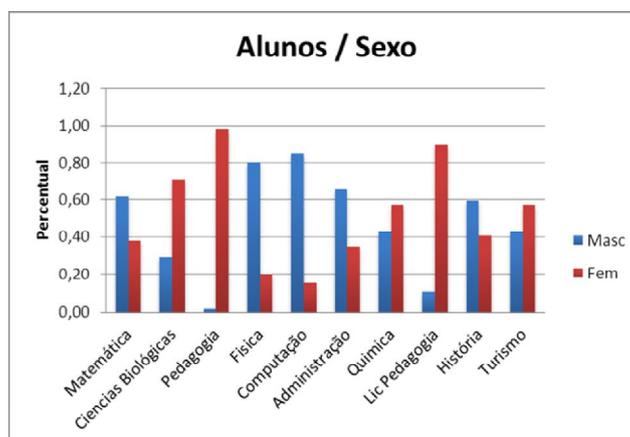


Figura 3 – Alunos da plataforma de acordo com o sexo

Com relação à “idade”, na maioria dos cursos nota-se predominância de alunos com idade superior a 28 anos. Chama atenção os cursos de Pedagogia, Licenciatura em

Pedagogia e História, onde essa predominância ultrapassa a marca de 70%. O curso de Química se apresenta como o mais diversificado com 50% de alunos com idade abaixo de 28 anos e 50% com idade acima de 28 anos (Figura 4).

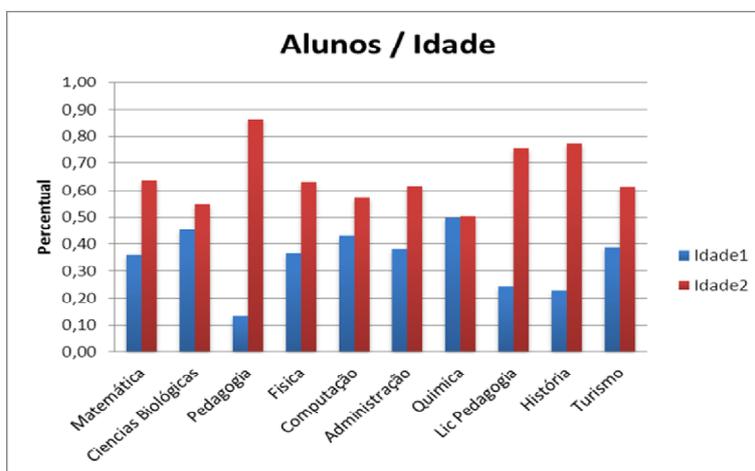


Figura 4 – Alunos da plataforma de acordo com a idade

Com relação ao “CR” a predominância é pela faixa Ruim (0,0 até 4,99). A exceção são os cursos de Pedagogia e Licenciatura em Pedagogia (Figura 5).

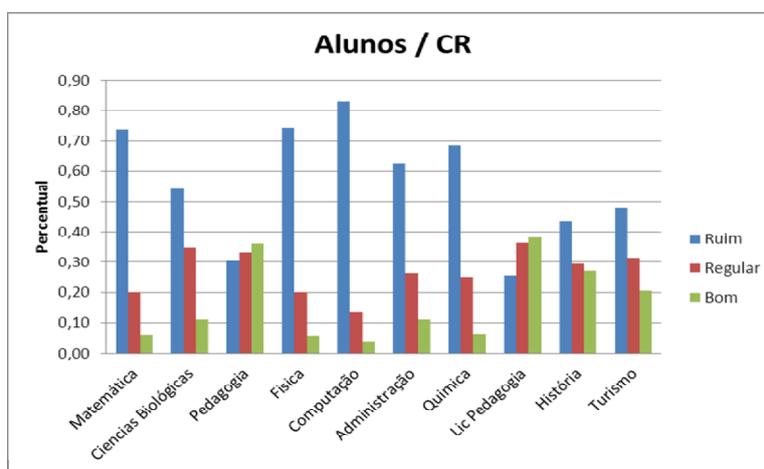


Figura 5 – Alunos da plataforma de acordo com o CR

O campo “Escola” apresenta a informação sobre de que tipo de escola o usuário é oriundo. Pode-se destacar que os alunos dos cursos da plataforma CEDERJ são oriundos, em sua maioria, de escolas Estaduais ou Particulares (Figura 6).

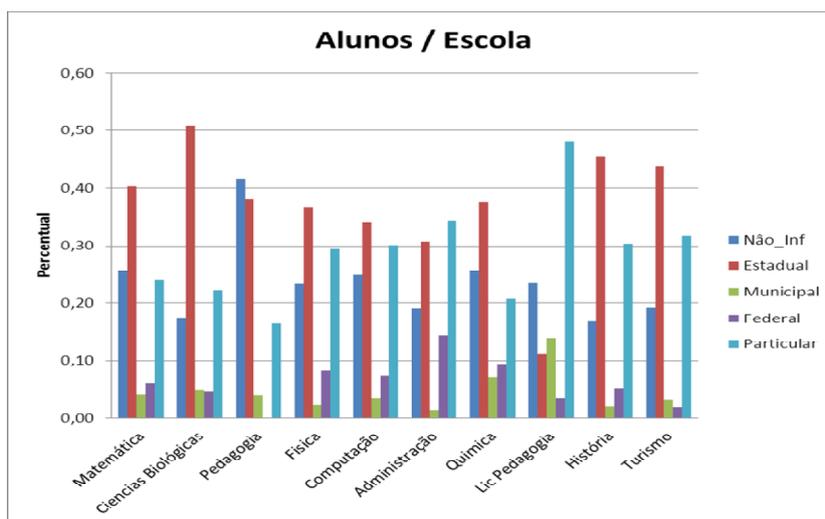


Figura 6 – Alunos da plataforma de acordo com a idade

De acordo com as características encontradas até o momento, pode-se traçar o padrão de usuário que utiliza a plataforma CEDERJ através da média das informações encontradas: 58% do sexo Feminino, 65% com idade acima de 28 anos, 54% tem CR Ruim e 52% são oriundos de escola Estadual.

Ao aplicar o algoritmo *Apriori* no *log* de acessos destes usuários, obteve-se as regras que indicam como se comportam os alunos que mais utilizam a plataforma CEDERJ (Figura 7).

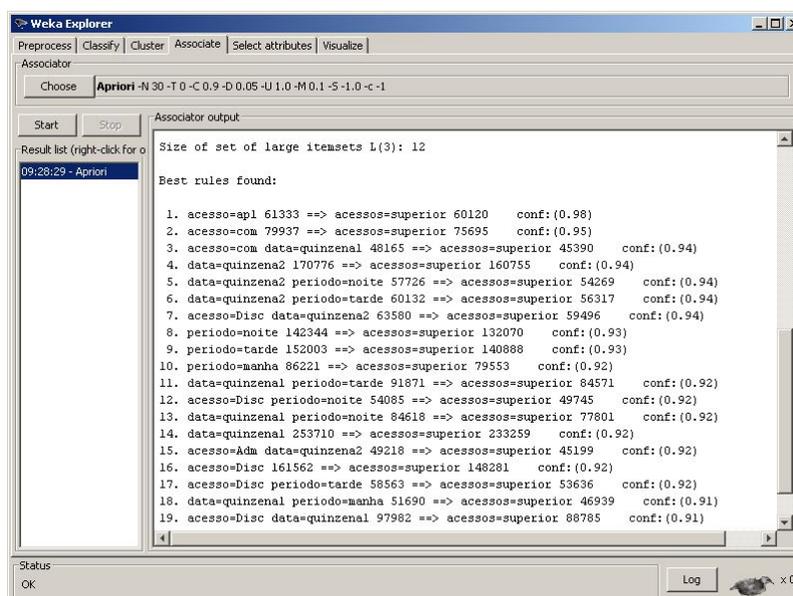


Figura 7 – Regras descobertas com o algoritmo Apriori

O algoritmo foi aplicado utilizando-se os parâmetros padrões da ferramenta WEKA, com exceção da quantidade de regras que foi aumentada de 10 para 30 regras a serem descobertas. A base de dados gerada possui 424.486 instâncias com um suporte mínimo de 10%. Foram descobertas 20 regras, dentre as quais pode-se destacar:

- 1) Todas as regras indicaram usuários cujo campo acesso foi indicado como “superior”, ou seja, realizam acima de 100 requisições a plataforma por mês;
- 2) A regra 3 aponta uma tendência que os usuários utilizam as ferramentas de comunicação, o fazem na primeira quinzena do mês. A regra 15 aponta que as ferramentas Administrativas são mais utilizadas na segunda quinzena do mês. Já as regras 7 e 19 indicam que as ferramentas de Disciplina são muito utilizadas durante todo o mês, não importando a data;

As regras não apontaram nenhuma forte tendência na utilização das ferramentas relacionadas a aplicativos. Em uma breve verificação, pode-se notar que tais ferramentas são pouquíssimas utilizadas na plataforma (14% de utilização).

Com relação ao período do dia, as regras apontam uma tendência a maior utilização da plataforma nos períodos da tarde e da noite. A única exceção é apontada pela regra 18 mostrando que na primeira quinzena do mês ocorreu uma tendência maior de acessos na parte da manhã.

3.8 Etapa 8: Utilização das Informações Obtidas

A última etapa é a utilização das informações obtidas no processo, incorporando o conhecimento em outro sistema para ações futuras ou simplesmente documentando e reportando as partes interessantes.

Após obter o padrão de usuários que utilizam o sistema plataforma através do processo de mineração, verifica-se a possibilidade de tais informações serem implementadas em um modelo matemático para geração de simulação, a fim de melhorar a qualidade no acesso ao sistema. Também seria possível mapear como esses usuários do padrão encontrado se comportam no sistema, identificando o caminho percorrido por eles quando estão “logados”. Com isso, seria possível responder questões como: Quais ferramentas o usuário frequentemente acessa após efetuar o “login”? Será que o usuário utiliza primeiramente as ferramentas de comunicação após o login ou será que utiliza as relacionadas a disciplinas? Existe diferenças de comportamento entre os usuários de diferentes cursos na utilização do sistema?

Para tal, haveria a necessidade de obter outras tendências significativas do comportamento do usuário, como: tempo de conexão, chegadas de requisições ao sistema, relação entre a navegação quando conectado ao sistema, entre outros. Também é importante realizar um levantamento dos usuários que não representam a maioria no sistema, como: usuários com CR Bom e menores que 28 anos. Torna-se necessário ainda verificar se esse comportamento é similar nos demais meses do ano letivo.

5. Conclusões

Neste trabalho foi apresentado a aplicação de técnicas de descoberta de conhecimento (KDD) com objetivo de identificar o padrão de usuários que melhor se adaptam a utilização de um sistema de Educação à Distância.

As técnicas de KDD foram aplicadas através de um estudo de caso único utilizando o *log* de acessos e dados cadastrais do CEDERJ/ Fundação CECIERJ para os

dez cursos oferecidos pela plataforma de ensino à distância. O método de pesquisa utilizado foi baseado nas etapas propostas por Fayyad [Fayyad 1996].

Os resultados das primeiras etapas demonstram que entre os cursos existem diferenças nos padrões de usuários que mais utilizam a plataforma. Por exemplo, nos cursos de Pedagogia e Licenciatura em Pedagogia existe a predominância do sexo feminino. Nos cursos de física e computação a predominância é a do sexo masculino. Nos demais cursos pode-se notar a utilização de ambos os sexos. Com relação à idade, em todos os cursos ocorre a predominância por usuários acima de 28 anos. O curso de pedagogia é o que apresenta a maior proporção de utilização por usuários nesta faixa de idade (87%). Com relação ao CR, a faixa predominante nos cursos é a Ruim (entre 0 e 4,9). A exceção ocorre no curso de Licenciatura em Pedagogia e Pedagogia, onde a proporção de usuários com CR Bom (5,0 e 6,99) e CR Regular (acima de 7,0) ultrapassam os de CR Ruim.

Um ponto com bastante divergência entre os cursos diz respeito à escola de onde o usuário é oriundo. Em grande parte dos cursos a escola estadual é apresentada como maioria.

O padrão apontado pelo estudo demonstra que a maioria dos usuários que utiliza o sistema é oriundo de escolas estaduais, possuem CR Ruim (entre 0 à 4,9), sexo Feminino e idade acima de 28 anos. Na etapa final, utilizou-se de regras de associação para descobrir como os usuários do padrão se adaptam e/ou se comportam ao utilizar o sistema. O algoritmo *Apriori* foi aplicado ao acesso (424486 instâncias) de todos os usuários do sistema que possuem o perfil padrão. O resultado mostrou que esses também são os usuários que mais enviam requisições/ações ao sistema (acima de 100).

As regras descobertas apontam para uma tendência maior de utilização das ferramentas disponibilizadas pela plataforma que dizem respeito às disciplinas, sendo as ferramentas relacionadas a aplicativos pouquíssimas utilizadas. Outra descoberta foi relacionada ao horário de maior utilização (tarde e noite). Com relação ao dia do mês, não existem períodos com maior ou menos utilização, no entanto, as regras apontam uma tendência maior de utilização das ferramentas relacionadas à comunicação na primeira quinzena do mês.

Referências

- Agrawal, R. and Srikant, R. (1994) "Fast Algorithms for Mining Association Rules", In Proceedings of the 20th international conference on very large databases, Santiago, Chile.
- Benicasa, A. and Paixão, R. (2006) "Mineração de dados como ferramenta para descoberta de conhecimento", Seminário de Informática – RS
- CEDERJ. Acessado em Fevereiro (2011), <http://portal.cederj.edu.br/fundacao/>.
- Dias, M. M. (2002), Parâmetros na escolha de técnicas e ferramentas de mineração de dados. Acta Scientiarum. Technology, Ed. EDUEM, v. 24, n.6, p. 1715-1725.
- Dilly, R. (1995), "Data Mining An Introduction", The Queen's University of Belfast, http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html, Acessado em 20/10/2011

- Fayyad, U. M. and Piatetsky-Shapiro and Smyth, P (1996), "From Data Mining to Knowledge Discovery: An Overview", In: *Advances in Knowledge Discovery and Data Mining*, AAAI Press.
- Garcia J. and Romero C., Ventura S. and Calders T. (2007), Drawbacks and solutions of applying association rules mining in learning management systems. *Proceedings of the International Workshop on Applying Data Mining in e-learning (ADML'07)*, Crete, Greece 2007, p. 13-22.
- Goebel, M.; Gruenwald, L. (1999), "A survey of data mining and knowledge discovery software tools", *ACM SIGKDD*, San Diego, v. 1, n. 1, p. 20-33.
- Nassar, Silvia M. and Wronski, Vilson R. and Ohira, Masanao et al. (2011), *SEstatNet - Sistema Especialista para o Ensino de Estatística na Web*. URL de acesso: <http://www.sestat.net> . Florianópolis - SC, Brasil. Acessado em 25/11/2011.
- Oliveira, A. G. and Garcia, D. F. (2005), *Mineração da Base de Dados de um Processo Seletivo Universitário*, Artigo, Centro Universitário de Formiga, Instituto de Ciências Sérias Aplicadas e Exatas, Curso de Ciência da Computação, Formiga, Minas Gerais.
- Romero, C. and Ventura, S. and Espejo P. G. and C. Hervas, (2008), Data mining algorithms to classify students. In: *Proceedings of International Conference on Educational Data Mining*, pp. 817.
- Santos, M. L. and Becker, K. (2002). *O uso da mineração de dados na web aplicado a um ambiente de ensino a distância*. Gramado. I Workshop de Teses e Dissertações em Banco de Dados.
- Silva M. P. S. (2011), *Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka*, Disponível em: www.sbc.org.br/bibliotecadigital/download.php?paper=35, Acesso em 02/10/2011.
- Silva, M., Boscaroli, C., Peres, S. (2003). *Análise de logs da Web por meio de técnicas de Data Mining*. Cascavel, Paraná, UNIOESTE – Universidade Estadual do Oeste do Paraná.
- Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA.: Sage, 1995.
- Witten, I.H. and E. Frank, L. Trigg, M. Hall, G. Holmes and S. J. Cunningham. (1999), "Weka: Practical machine learning tools and techniques with Java implementations," in H. Kasabov and K. Ko, eds., *ICONIP/ANZIIS/ANNES'99 International Workshop*, Dunedin.
- Wu, X.; Kumar, V.; Ross Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; Mclachlan, G. J.; Ng, A.; Liu, B.; Yu, P. S.; Zhou, Z.; Steinbach, M.; Hand, D. J.; Steinberg, D. Top 10 algorithms in data mining. *Knowledge Information Systems*, v. 14, n. 1, p. 1-37, 2007. doi:10.1007/s10115-007-0114-2