Revista Brasileira de Informática na Educação, Volume 19, Número 2, 2011

Mineração de Dados Educacionais: Oportunidades para o Brasil

Ryan Shaun Joazeiro de Baker
Department of Social Sciences and Policy Studies
Worcester Polytechnic Institute
100 Institute Road, Worcester, MA 01609 USA
rsbaker@wpi.edu

Seiji Isotani
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA
sisotani@cs.cmu.edu

Adriana Maria Joazeiro Baker de Carvalho
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA
carvalho@cs.cmu.edu

Resumo *A mineração de dados educacionais (EDM) é uma área recente de pesquisa que tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Atualmente ela vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino. Apesar dos esforços de pesquisadores brasileiros, essa área ainda é pouco explorada no país. Para divulgar alguns dos resultados desta área este artigo apresenta uma revisão das pesquisas realizadas na área, dando ênfase aos métodos e aplicações que vêem influenciado, com sucesso, a pesquisa e a prática da educação em vários países. Serão discutidas as condições que viabilizam a pesquisa da EDM no cenário internacional e quais os desafios para consolidar a área no Brasil. Além disso, também será abordado o potencial impacto da EDM na melhora da qualidade dos cursos na modalidade educação a distância (EAD) que vêm recebendo incentivo governamental e um crescente número de alunos matriculados.*

Palavras-Chave: *Mineração de Dados Educacionais, Educação a Distância*

Abstract *Educational Data Mining (EDM) is the research area concerned with the development and use of data mining methods for exploring data sets collected in educational settings. In recent years, EDM has become established internationally as a field and research community, with evidence of considerable potential to improve the quality of education. Though there have been efforts to establish EDM research in Brazil, EDM is not yet well established in Brazil. Towards increasing awareness of EDM research in Brazil, this paper presents a review of research on EDM, discussing methods and successful applications of EDM research which have influenced research and educational practice internationally. The article discusses some of the enabling conditions for EDM research, and the challenges that must be met for this field to reach its full potential in Brazil. In specific, we discuss the potential that EDM research has to benefit the increasing number of Brazilian distance learners.*

Keywords: *Educational Data Mining, Distance Learning*

Recebido: 10 de Maio de 2011 / Aceito: 13 de Agosto de 2011 / Publicado: 24 de Agosto de 2011
DOI: 10.5753/RBIE.2011.19.02.03

1 Mineração de Dados Educacionais

O termo **Mineração de dados**, também conhecido como *Descoberta de Conhecimentos em Bancos de Dados*, ou KDD (do inglês, "*Knowledge Discovery in Databases*"), refere-se a disciplina que tem como objetivo descobrir "novas" informações através da análise de grandes quantidades de dados [41]. O termo "novas informações" refere-se ao processo de identificar relações entre dados que podem produzir novos conhecimentos e gerar novas descobertas científicas.

As informações sobre a relação entre dados e, posteriormente a descoberta de novos conhecimentos, podem ser muito úteis para realizar atividades de tomada de decisão. Por exemplo, ao minerar os dados de um estoque de supermercado poderia-se descobrir que todas as sextas-feiras uma marca específica de cerveja se esgota nas prateleiras e, portanto, um gerente que obtém esta "nova informação" poderia planejar o estoque do supermercado para aumentar a quantidade de cervejas desta marca as sextas-feiras. Analogamente, é possível minerar dados de alunos para verificar a relação entre uma abordagem pedagógica e o aprendizado do aluno. Através desta informação o professor poderia compreender se sua abordagem realmente está ajudando o aluno e desenvolver novos métodos de ensino mais eficazes. A Mineração de dados tem sido aplicada em diversas áreas do conhecimento, como por exemplo, vendas, bioinformátíca, e ações contra-terrorismo. Recentemente, com a expansão dos cursos a distância e também daqueles com suporte computacional, muitos pesquisadores da área de Informática na Educação (em particular, Inteligência Artificial Aplicada à Educação) têm mostrado interesse em utilizar mineração de dados para investigar perguntas científicas na área de educação (e.g. quais são os fatores que afetam a aprendizagem? Ou como desenvolver sistemas educacionais mais eficazes?). Dentro deste contexto, surgiu uma nova área de pesquisa conhecida como "**Mineração de Dados Educacionais**" (do inglês, "*Educational Data Mining*", ou EDM). A EDM é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Assim, é possível compreender de forma mais eficaz e adequada os alunos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. Por exemplo, é possível identificar em que situação um tipo de abordagem instrucional (e.g. aprendizagem individual ou colaborativa) proporciona melhores benefícios educacionais ao aluno. Também é possível verificar se o aluno está desmotivado ou confuso e, assim, personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem.

A comunidade de EDM vem crescendo rapidamente. Em 2008 criou-se a Conferência Internacional sobre Mineração de Dados Educacionais (*International Conference on Educational Data Mining*), após uma sequência de worshops bem sucedidos realizados anualmente desde 2004. Em sua terceira edição, foram submetidos 74 artigos originais para esta conferência e o número de participantes aumentou consideravelmente em relação aos anos anteriores. Criou-se também a Revista de Mineração de Dados Educacionais (*Journal of Educational Data Mining*), que publicou seu primeiro volume em Novembro de 2009. Além da consolidação da conferência e da revista na área de EDM, a comunidade também publicou dois livros sobre o assunto em 2006 e 2010 (*Data Mining in e-learning e Handbook of Educational Data Mining*).

Contudo, no Brasil ainda são poucos os trabalhos publicados nesta área de pesquisa. Um dos trabalhos pioneiros no uso de mineração de dados na educação foi publicada por Brandão et al. [13] analisando dados do programa nacional de informática na educação. Um outro trabalho pioneiro no Brasil que analisou dados da avaliação de alunos é apresentada por Pimentel e Omar [35]. Com o objetivo de divulgar esta área no Brasil, este artigo apresenta uma breve introdução de alguns métodos e aplicações da EDM e a visão dos autores sobre o potencial benéfico que a EDM pode trazer ao sistema educacional brasileiro, principalmente para a educação a distância.

2. Métodos para EDM

Existem muitos métodos utilizados em EDM que são originalmente da área de mineração de dados [41]. Contudo, de acordo com Baker [4], muitas vezes estes métodos precisam ser modificados, por causa da necessidade de considerar a hierarquia (em diversos níveis) da informação. Além disso, existe uma falta de independência estatística nos tipos de dados encontrados ao coletar informações em ambientes educacionais. Por causa disso, diversos algoritmos e ferramentas utilizadas na área de mineração de dados não podem ser aplicadas para analisar dados educacionais sem modificação. Em particular, ferramentas importantes de mineração de dados, como por exemplo Weka [cf. 41], não oferecem apoio para validação cruzada entre os dados no nível do aluno ou da classe. A validação cruzada permite verificar a corretude de um modelo gerado a partir da análise de dados de treinamento (*training data*). Essa validação oferece uma estimativa de como o modelo irá se comportar ao analisar um conjunto novo dados. Validação cruzada ao nível de aluno ou classe é fundamental em dados educacionais, pois existe uma grande quantidade de dados por aluno e as conclusões obtidas ao utilizar métodos de mineração de dados precisam garantir que o modelo encontrado possa ser utilizados para inferir o comportamento ou a aprendizagem novos alunos e/ou classe. RapidMiner [cf.

30] é uma ferramenta que oferece um melhor apoio a estas análises, embora ainda exija do usuário um grande esforço para obter a validação desejada. Devido a esta lacuna na área de mineração de dados, muitos pesquisadores que publicam na área de EDM utilizam modelos desenvolvidos na área de psicometria [e.g. 10, 18, 33].

Existem várias linhas de pesquisa na área de EDM. Muitas delas derivadas diretamente da área de mineração de dados. Assim, nos parágrafos a seguir faremos uma breve introdução de alguns dos tópicos mais interessantes da área.

Uma taxonomia das principais sub-áreas de pesquisa em EDM é apresentada em [4]:

- Predição (*Prediction*)
  - Classificação (*Classification*)
  - Regressão (*Regression*)
  - Estimação de Densidade (*Density Estimation*)
- Agrupamento (Clustering)
- Mineração de relações (*Relationship Mining*)
  - Mineração de Regras de associação (*Association Rule Mining*)
  - Mineração de Correlações (*Correlation Mining*)
  - Mineração de Padrões Sequenciais (*Sequential Pattern Mining*)
  - Mineração de Causas (*Causal Mining*)
- Destilação de dados para facilitar decisões humanas (*Distillation of Data for Human Judgment*)
- Descobertas com modelos (*Discovery with Models*)

As três primeiras categorias dessa taxonomia são de interesse tanto da área de EDM quanto da área de mineração de dados em geral. As sub–categorias de Predição: Classificação, Regressão e Estimação de Densidade estão diretamente relacionadas as categorias dos métodos de mineração de dados apresentados por Moore [32].

Na área de **predição**, a meta é desenvolver modelos que deduzam aspectos específicos dos dados, conhecidos como variáveis preditivas (*predicted variables*), através da análise e fusão dos diversos aspectos encontrados nos dados, chamados de variáveis preditoras (*predictor variables*). A Predição necessita que uma certa quantidade dos dados seja manualmente codificada para viabilizar a correta identificação de uma ou mais variáveis preditoras previamente conhecidas (a codificação e a identificação das variáveis não precisam ser perfeitas). Como indicado na taxonomia, existem três tipos de predição: classificação, regressão, e estimação de densidade. A estimação de densidade é raramente utilizada na EDM devido a falta de independência estatística dos dados. Em classificação, a variável preditora é binária ou categórica. Alguns algoritmos populares na EDM, disponíveis em ferramentas como o RapidMiner [30], incluem árvores de decisão, regressão logística (para predições binárias), e regressão *step*. Quando a variável preditora é um número, os algoritmos de regressão mais populares incluem regressão linear, redes neurais, e máquinas de suporte vetorial. Para classificação e regressão, as variáveis preditoras podem ser categóricas ou numéricas; métodos diferentes ficam mais (ou menos) efetivos, dependendo das características das variáveis preditoras utilizadas.

Existem dois benefícios de se utilizar métodos de predição em EDM. Primeiro, métodos de predição são utilizados para estudar quais aspectos de um modelo são importantes para predição, dando informação sobre o construto sendo examinado (exemplos de constructo modelado incluem curvas de aprendizagem e representações de tipos variados de comportamento). Esta estratégia é frequentemente utilizada em pesquisas que tentam, de forma direta, predizer os benefícios educacionais para um conjunto de estudantes [e.g. 37], sem primeiro predizer os fatores mediantes ou intermediários. Ou seja, o objetivo é verificar o quanto o aluno aprender sem considerar as diversas variáveis que influenciam a aprendizagem como, por exemplo, variáveis relacionadas ao comportamento do estudante [33]. Segundo, os métodos de predição auxiliam a predizer o valor das variáveis utilizadas em um modelo. Essa abordagem é necessária, pois analisar *todos* os dados de um grande banco de dados para gerar um modelo é tipicamente financeiramente inviável, além de consumir muito tempo [7]. Assim, o modelo pode ser construído utilizando parte dos dados e então ser aplicado para modelar dados mais extensos [6]. Esse tipo de técnica pode auxiliar no desenvolvimento e uso de atividades instrucionais, pois consegue estimar os benefícios educacionais antes mesmo da atividade ser aplicada com os alunos.

Na área de **agrupamento**, o objetivo principal é achar dados que se agrupam naturalmente, classificando os dados em diferentes grupos e/ou categorias. Estes grupos e categorias não são conhecidos inicialmente. Através de técnicas de agrupamento os grupos/categorias são automaticamente identificados através da manipulação das características dos dados. É possível criar esses grupos/categorias utilizando diferentes unidades de análise, por exemplo é possível achar grupos de escolas (para investigar as diferenças e similaridades entre escolas), ou achar grupos de alunos (para investigar as diferenças e similaridades entre alunos), ou até grupos de atos (para investigar padrões de comportamento) [2].

Em **mineração de relações**, a meta é descobrir possíveis relações entre variáveis em bancos de dados. Esta tarefa pode envolver a tentativa de aprender quais variáveis são mais fortemente associadas com uma variável específica, previamente conhecida e importante, ou pode envolver as relações entre quaisquer variáveis presentes

nos dados. Para identificar essas relações, existem quatro tipos de mineração: (a) regras de associação; (b) correlações; (c) sequências; ou (d) causas.

Na **mineração de regras de associação**, procura-se gerar/identificar regras do tipo *se-então* (*if-then*) que permitam associar o valor observado de uma variável ao valor de uma outra variável. Ou seja, caso uma condição seja verdadeira (e.g. variável *Y* possui valor 1) e uma regra associe essa condição ao valor de uma outra variável *X*, então podemos inferir o valor desta variável *X*. Por exemplo, ao analisar um conjunto de dados seria possível identificar uma regra que faz a associação entre a variável "*objetivo do aluno*", uma variável binária que pode ter os valores *alcançado* ou *não alcançado*, e uma outra variável binária "*pedir ajudar ao professor*" que pode ter os valores *sim* ou *não*. Neste contexto, **se** o aluno tem como objetivo aprender geometria, mas está com dificuldade (i.e. a variável *objetivo do aluno* tem valor *não alcançado*), **então** é provável que ele peça ajuda do professor (i.e. a variável *pedir ajuda ao professor* tem valor positivo).

Em **mineração de correlações**, a meta é achar correlações lineares (positivas ou negativas) entre variáveis. Por exemplo, ao analisar um conjunto de dados, seria possível identificar a existência de uma correção positiva entre uma variável que indica a quantidade de tempo que um aluno passa externalizando comportamentos que não estão relacionados as tarefas passadas pelo professor (e.g. conversas paralelas, brincadeiras e outras perturbações que ocorrem em sala de aula) e a nota que este aluno recebe na próxima prova.

Em **mineração de sequências**, o objetivo principal é achar a associação temporal entre eventos e o impacto destes eventos no valor de uma variável. Neste caso, é possível determinar qual trajetória de atos e ações de um aluno pode, eventualmente, levar a uma aprendizagem efetiva. Dessa forma, é possível criar um conjunto de atividades instrucionais que podem melhorar a qualidade do ensino fazendo com que os alunos externalizem ações que vão ajudá-los a construir seu conhecimento e desenvolver as habilidades necessárias para trabalhar com o conteúdo apresentado pelo professor.

Em **mineração de causas**, desenvolve-se algoritmos e técnicas para verificar se um evento causa outro evento através da analise dos padrões de covariância (uma sistema que faz isso é TETRAD [39]). Por exemplo, se considerarmos o exemplo anterior onde um aluno externaliza comportamentos inadequados que não contribuem para resolver a tarefa dada pelo professor. Nesta situação o aluno, em muitos casos, recebe uma nota ruim na prova final. Nesta situação, o comportamento do aluno pode ser a causa dele não aprender e, assim, resultado em uma performance ruim na prova. Contudo, pode ser que o aluno externalize tal comportamento inadequado devido a dificuldade em aprender, e portanto, a causa da performance ruim na prova não é o comportamento em sí, mas sim a dificuldade de aprendizagem do aluno. Aalisando o padrão de covariância, a mineração de causa pode inferir qual evento foi a causa do outro.

Na área de **destilação de dados para facilitar decisões humanas**, são realizadas pesquisas que tem como objetivo apresentar dados complexos de forma a facilitar sua compreensão e expor suas características mais importantes. Através da destilação é possível que os dados sejam utilizados por pessoas para inferir aspectos sobre os dados e, assim, tomar decisões que anteriormente não poderiam ser tomadas e nem automatizadas apenas com o uso dos métodos da EDM. Os métodos dessa sub-área da EDM facilitam a visualização da informação contida nos dados educacionais coletados por softwares educacionais [22, 25]. Estes métodos "purificam" os dados para auxiliar as pessoas na identificação de padrões. Em diversas ocasiões, esses padrões são previamente conhecidos, mas são difíceis de serem visualizados e/ou descritos formalmente. Por exemplo, uma visualização clássica em EDM é a ***curva de aprendizagem***. Essa curva indica o nível de aprendizagem de um aluno (ou de um conjunto de alunos) ao longo do tempo. Ela é apresentada num plano cartesiano conforme mostra a figura 1. Nesta curva relaciona-se o número de oportunidades que o aluno praticou um componente de conhecimento[1] (apresentado no eixo *x*) e a sua performance (porcentagem de valores corretos, apresentada no eixo *y*). Uma curva que desce rapidamente no inicio do gráfico e depois gradativamente diminui sua inclinação indica que o modelo de conhecimento é bem especificado. Ou seja, o modelo representa corretamente quais as relações entre os componentes de conhecimento e as atividades realizadas pelos alunos. Essas atividades oferecem a oportunidade de praticar os componentes de conhecimento relacionados e ao decorrer deste processo o aluno aprende a medida que suas habilidades e conhecimentos são testados.

Caso a curva de aprendizagem possua diversos pontos fora dos locais esperados, ou seja, porcentagem de erros muito acima ou muito abaixo do esperado dado o número de oportunidades, isso indica que o modelo utilizado não está bem refinado e provavelmente mais de um componente de conhecimento está sendo tratado no mesmo problema [16]. No caso da Figura 1 a curva em vermelho representa dados de alunos e a curva tracejada em azul representa a curva esperada calculada utilizando algoritmos de predição implementados na plataforma *Datashop*[2] [27]. Observe que apesar de alguns pontos estarem um pouco acima ou abaixo do esperado a curva em vermelho desce gradativamente seguindo a curva esperada. Ou seja,

[1] Um componente de conhecimento pode ser definido como um conceito, uma habilidade, uma regra ou um princípio utilizado para resolver uma tarefa. Maiores informações sobre a definição de componentes de conhecimento podem ser obtidas em [27].

[2] Repositório para armazenamento e análise de dados educacionais https://pslcdatashop.web.cmu.edu/

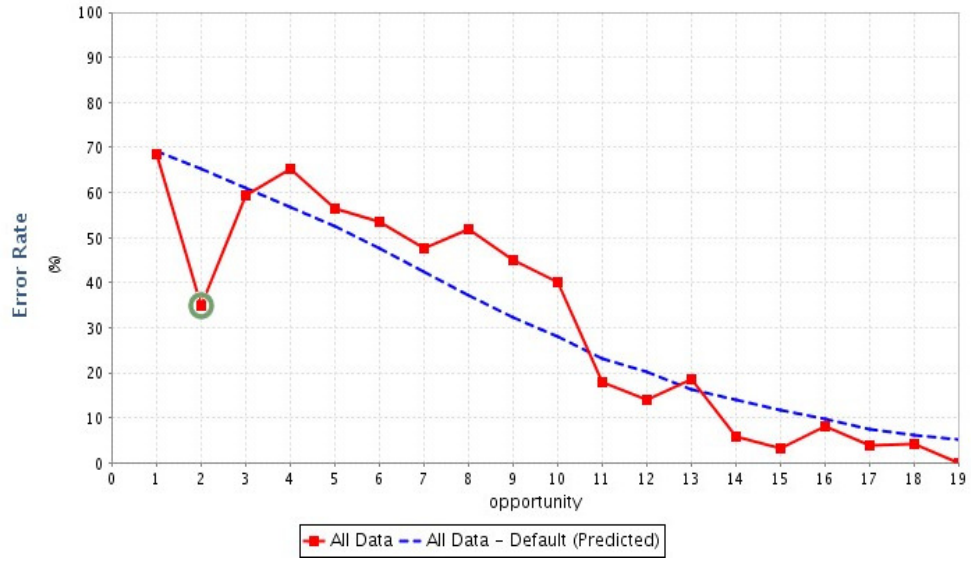

Figura 1. Curva de aprendizagem utilizada na plataforma Datashop. A curva em vermelho representa os dados obtidos pelos alunos e a curva tracejada em azul representa a curva esperada, de acordo com o modelo desenvolvido.

o modelo possui algumas falhas, mas corretamente indica que os alunos aprenderam ao longo do curso o conteúdo e as habilidades (componentes de conhecimento) desejadas. Observe que esse modelo pode ser utilizado para identificar a evolução da aprendizagem de qualquer aluno. Dependendo das interação entre o aluno e as atividades, é possível verificar o quanto o aluno aprendeu ou estimar o quanto ele irá aprender após um conjunto de atividades realizadas.

O uso da destilação de dados também é muito útil para categorizar as ações dos estudantes. Através desta categorização é possível auxiliar o desenvolvimento de um modelo de predição mais robusto [8].

3. Disponibilização dos Dados: Uma Condição Essencial

Atualmente, as tendências na área indicam um grande crescimento das pesquisas em EDM no cenário internacional, em particular nos Estados Unidos, Canadá, e Espanha. Este grande crescimento é resultado direto de dois outros fatores que discutiremos a seguir. Primeiro, a difusão e utilização de software educacionais que produzem grandes quantidades de dados educacionais bem estruturados. Por exemplo, o *Cognitive Tutor*[1] é um tipo de sistema tutor inteligente que produz quantidades significativas de dados de boa qualidade. Este sistema é utilizado anualmente por mais de 500 mil alunos em, aproximadamente, 2000 escolas espalhadas pelos Estados Unidos. Dados vindos do *Cognitive Tutor*, e também de outros sistemas educacionais (e.g. MathTutor[2]), estão disponíveis gratuitamente para qualquer pesquisador, através de repositórios educacionais como o *DataShop* criado pelo Centro de Ciências da Aprendizagem de Pittsburgh (PSLC - *Pittsburgh Science of Learning Center*). Os dados disponíveis no *Datashop* estão sendo utilizados por mais que 400 pesquisadores em todo o mundo. Segundo Baker & Yacef [9], dados de alunos retirados do *DataShop* foram utilizados em 14% dos artigos publicados na Conferência Internacional sobre Mineração de Dados Educacionais em 2008 e 2009.

O segundo fator que vem promovendo o crescimento da EDM é o uso de sistemas computacionais de gerenciamento de curso/aprendizagem (e.g. LMS – *learning management systems*; e CMS – *content management systems*) como o Moodle e o WebCat. Estes sistemas vêm sendo adotados por muitos professores, escolas e universidades em todo mundo. Além disso, já existem softwares que permitem que pesquisadores utilizem os dados gerados por estes sistemas de forma que seja possível realizar a mineração de dados [1, 38]. Nos Estados Unidos, dados de escolas e distritos (conjunto de escolas de uma cidade) estão começando a ser disponibilizados aos pesquisadores através de bancos de dados públicos como, por exemplo, o banco de dados do Centro Nacional de Estatísticas Educacionais (*National Center for Education Statistics*). Estes recursos permitem que pesquisadores possam, mais facilmente, obter grandes quantidades de dados reais e relevantes para realizar análises utilizando técnicas provindas da área de EDM. Os pesquisadores que fazem uso

[1] http://www.carnegielearning.com/

[2] https://mathtutor.web.cmu.edu/

destes dados podem conduzir pesquisas com alta validade ecológica, ou seja, os resultados podem ser utilizados no contexto escolar, enquanto que ao mesmo tempo evita-se os custos tradicionais da pesquisa e da coleta de dados nesta área.

Com a difusão destes repositórios de dados educacionais abertos diminui-se a necessidade de (a) recrutar escolas, professores, e estudantes; (b) realizar estudos convencionais que requerem recursos humanos especializados; (c) ir para escolas e conduzir experimentos que duram dias ou até semanas; (d) inserir, formatar e digitalizar os dados obtidos; e etc. Essa abordagem poderá salvar grande parte do tempo e dos custos envolvidos em pesquisas educacionais. Além disso, os resultados poderão ser obtidos mais rapidamente, serão mais precisos e, finalmente, proporcionarão o desenvolvimento de práticas pedagógicas que podem ser utilizados para melhorar a qualidade do ensino de forma eficaz.

4. Principais Aplicações da EDM

As pesquisas em mineração de dados educacionais vêm oferecendo contribuições significativas para a teoria e a prática da educação [9]. Podemos citar diversos exemplo do uso de métodos da EDM para melhorar os modelos de conhecimento do estudante em vários diferentes domínios como ensino de língua estrangeira, geometria, química, física e muitos outros [10, 14, 17, 33]. Um dos benefícios desse avanço foi a redução considerável do tempo gasto pelos alunos para desenvolver suas habilidades acadêmicas, principalmente em domínios como a matemática [15].

Métodos da EDM também viabilizaram a expansão do conhecimento científico relacionado aos estados emocionais do aluno (e.g. motivado, frustrado, confuso, etc). Eles também têm auxiliado a identificar a relação entre estes estados emocionais e o comportamento apresentado pelo aluno, principalmente quando ocorre a externalização intencional de ações inadequadas (trapaças) ao interagir com software educacionais (em inglês, esse comportamento inadequado é conhecido como "*gaming the system*"). Com o uso dos métodos da EDM (e.g. mineração de causas e correlações) em conjunto com softwares educacionais é 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 [8]. Através desta verificação a área da 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.

Pesquisas nessa área também proporcionaram modelos automatizados que podem ser utilizados durante a interação dos alunos com os programas educacionais para identificar quando os alunos estão tentando trapacear para conseguir melhores notas sem ter aprendido o conteúdo adequadamente [6]. Diversos algoritmos que analisam em tempo real os dados das interações dos alunos com a interface do sistema foram desenvolvidos para verificar automaticamente quando comportamentos inadequados ocorrem. Essa funcionalidade permite que sistemas educacionais apresentem comportamentos "inteligentes" oferecendo suporte e *feedback* apropriados para melhorar a qualidade da aprendizagem dos alunos.

Um exemplo deste tipo de sistema inteligente é apresentado por Baker e colegas [5]. Neste trabalho, os autores desenvolveram um personagem (Scooter) que reage de acordo com o comportamento apresentado pelo aluno. Quando o aluno interage com o sistema de forma adequada, Scooter faz sinal de positivo conforme mostra a imagem no canto superior esquerdo da Figura 2. Quando o aluno tenta trapacear, por exemplo, pedindo ajuda ao sistema diversas vezes para tentar obter a resposta final de um exercício sem ao menos tentar resolvê-lo, então Scooter muda seu comportamento conforme mostra a imagem no canto inferior esquerdo da Figura 2. Como nesta situação o sistema não possui dados suficientes para determinar se o aluno realmente aprendeu ou não o conteúdo desejado então Scooter tenta diagnosticar o conhecimento do aluno através de uma sequência perguntas adicionais (à direita da Figura 2) que possuem dois propósitos: (1) verificar se o aluno aprendeu o conteúdo corretamente; (2) revisar o conteúdo para auxiliar aqueles alunos que não aprenderam corretamente. Como resultado, os autores do trabalho enfatizam que o comportamento deste personagem auxiliou o professor a identificar os alunos que não estavam aprendendo corretamente e também incentivou os alunos a manter um comportamento adequado para aprender de forma eficaz o conteúdo da matéria.

Além da contribuição para o desenvolvimento de programas educacionais eficazes, resultados da EDM também influenciaram áreas mais tradicionais da educação. Um resultado importante foi apresentado por Beck e Mostow [11] que, através da análises de dados de atividades relacionadas a leitura, demonstrou que re-ler a mesma história é vantajoso para crianças com desabilidades (que lêem vagarosamente), mas não oferecer benefícios para as outras crianças que estão aprendendo a ler. Nesse último caso ler histórias diferentes proporcionam mais benefícios à aprendizagem [11]. Esse resultado também indica que existe a necessidade de se oferecer suporte para personalizar a forma de apresentar o conteúdo em classe auxiliando os alunos a atingirem os objetivos desejados.

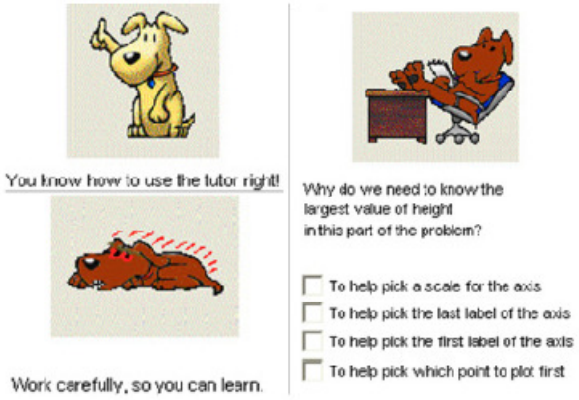

Figura 2. Scooter, um personagem que adapta suas ações a partir da análise das interações realizadas pelo aluno.

Embora a EDM seja uma área de pesquisa ainda recente, artigos dessa área são freqüentemente citados pela comunidade de Computação aplicada à Educação. Em seu artigo sobre o estado da arte na área de EDM, Baker e Yacef apresentam a lista dos artigos mais citados na área até o momento de acordo com o *google scholar* [9]. Destaca-se que muitos artigos tiveram um grande número de citações em apenas alguns meses após sua publicação. Por exemplo, em apenas 7 meses (entre o momento da publicação e a escrita do artigo sobre o estado da arte realizado por Baker e Yacef), 5 artigos publicados em conferências e revistas relacionadas à EDM tiveram mais de 80 citações adicionais.

5. EDM: Oportunidades pelo Brasil

A área de EDM vem estabelecendo uma comunidade forte tanto nos EUA quanto na Europa. Contudo, no Brasil a comunidade e as pesquisas nessa área ainda estão em seu estágio inicial. Nas seis edições de workshops e conferências internacionais em mineração de dados educacionais, realizados desde 2004, só um artigo publicado teve a co-autoria de um pesquisador afiliado a uma instituição Brasileira (e este artigo envolveu dados coletados nos Estados Unidos) [34]. Na JEDM (*Journal of Educational Data Mining*), a revista mais importante da área, não existe até o momento nenhum artigo com autores brasileiros.

Ao mesmo tempo, o potencial para a pesquisa, o desenvolvimento e a aplicação da EDM em ambientes educacionais vem crescendo muito no Brasil. Em particular, com a criação da Universidade Aberta do Brasil e a legalização de diversos cursos na modalidade à distância, criou-se grandes oportunidade para pesquisas em EDM no país. Atualmente, o Brasil é um dos países que mais cresce no número de cursos oferecidos na modalidade Educação a Distancia (EaD). Em 2008, mais de 750.000 estudantes brasileiros participaram de cursos e/ou programas de EaD pela internet [31]. Esta mudança no sistema educacional brasileiro proporciona um ambiente singular na qual a mineração de dados pode proporcionar impactos muito positivos.

Os dados obtidos em ambientes de EaD, como grandes quantidades de texto envolvendo discussões (síncronas e assíncronas) em chats, fóruns de discussão, wikis, blogs, e outras formas de interação textual entre estudante-estudante e estudante-professor, têm excelente potencial de serem utilizadas para realizar-se mineração de texto de descobrir modelos interessantes sobre os alunos. Através das diversas técnicas da EDM, brevemente apresentadas na Seção 2, é possível averiguar se as diversas ferramentas disponíveis em ambientes online são realmente eficazes para auxiliar a aprendizagem do aluno. Por exemplo, apesar de muitos ambientes virtuais de aprendizagem possuírem ferramenta de *chat* (sala de bate papo), são poucos os trabalhos científicos que analisam os dados obtidos por essa ferramenta e correlacionam o conteúdo das conversas com a aprendizagem dos alunos [3]. O mesmo problema ocorre com fóruns de discussão [21]. Perguntas como, quais foram os tópicos discutidos? Quais as conclusões alcançadas pelos alunos? Quem interagiu com quem? Qual a porcentagem de mensagens relacionadas ao assunto da aulas? São algumas das perguntas que precisam ser respondidas para que o professor, ou o próprio software educacional, possa compreender quais processos de interação facilitam a aprendizagem e quais destes dificultam o desenvolvimento do conhecimento do aluno.

A pesquisa de Prata et al [34], mostra um exemplo da aplicação da EDM em atividades colaborativas realizadas em ambientes de EaD. Nesta pesquisa, os autores estudaram a relação entre os atos de colaboração e a aprendizagem dos alunos em escolas do ensino médio. Os resultados obtidos indicam que os alunos que aprendem bem o conteúdo apresentado pelo professor tem maior chance de mostrar comportamentos inapropriados específicos (insultos) durante o andamento do curso. Este resultado é interessante, pois muitos professores acreditam que os alunos com baixa performance escolar são os maiores responsáveis por atrapalhar a aprendizagem dos outros alunos, contradizendo os dados obtidos por Prata et al. Essas visões antagônicas sugerem que os comportamentos dos alunos precisam ser melhor estudados e compreendidos para identificar a razão de sua ocorrência. No trabalho de Isotani et al. [23] identificou-se que muitos dos comportamentos inadequados que surgem durante o uso de ambientes colaborativos ocorrem devido a falta de um planejamento adequado das atividades colaborativas. Estudando tais comportamentos é possível identificar as características daqueles alunos que conseguem aprender através da EaD e também daqueles que não conseguem.

Para realizar estes estudos de comportamento, ferramentas como o TagHelper [20] utilizadas por Prata e colegas para analisar atividades colaborativas podem facilitar e agilizar o trabalho de pesquisadores e educadores. Assim, será possível compreender mais rapidamente o impacto dos comportamentos e das interações entre alunos e professores no processo de ensino-aprendizagem em ambientes presenciais e de EaD. Como conseqüência direta, técnicas mais efetivas serão desenvolvidas para ajudar professores a criarem abordagem pedagógicas e ambientes computacionais que incentivam a aprendizagem aumentando as chances dos alunos aprenderem de maneira mais rápida e eficaz..

No Brasil, o desafio de analisar e compreender o comportamento dos alunos é muito grande devido a diversidade da população. De acordo com Blanchard et al. [12] existe uma correlação entre os dados sócio-culturais dos alunos e suas ações, atitudes e comportamentos apresentados durante a aprendizagem. Isso significa que para desenvolver ambientes de EaD efetivos no Brasil, onde a diversidade cultural e econômica é grande, será necessário o desenvolvimento de algoritmos e ferramentas computacionais que levam em consideração a realidade brasileira. Pesquisas nessa área auxiliarão a descobrir formas inteligentes de difundir e personalizar o conteúdo do cursos para apoiar o aluno de acordo com sua personalidade, religião, raça, cultura, idade, sexo, e etc, fazendo com que cada indivíduo tenha uma experiência única dentro do ambiente virtual de aprendizagem. Tal possibilidade é uma das grande vantagens da EaD e tem sido defendida ao longo dos anos por muitos educadores. A quase 15 anos atrás Preti identificou a EaD como essencial para desenvolvimento da educação no Brasil, dizendo que a EaD é "*uma modalidade de se fazer educação, onde se democratiza o conhecimento*" [36].

Para que a EaD e a EDM tenham impacto na sociedade brasileira é necessário que pesquisadores e educadores comecem a utilizar os dados obtidos em ambientes de EaD de forma estruturada e com objetivos bem definidos. Recentemente, alguns trabalhos publicados no Simpósio Brasileiro de Informática na Educação tiveram como tema principal o uso da EDM para analisar textos, apoiar a produção de conteúdo educacional, apoiar a aprendizagem em ambientes virtuais de aprendizagem e criar serviços semânticos [28].

Uma pesquisa brasileira que merece destaque nessa área é apresentada por Kampff [24]. Em sua tese de doutorado , Kampff utiliza técnicas da área de EDM para identificar comportamentos e características de alunos com alto risto de evazão ou reprovação em ambientes virtuais de aprendizagem. Ao verificar que um aluno possui tais comportamentos/características o sistema alerta o professor que poderá tomar as decisões pedagógicas necessárias para que o aluno fique mais motivado, volte a aprender e não desista do curso. Esse recurso é muito interessante, pois o professor pode melhorar suas técnicas de ensino, e verificar quais alunos estão passando por dificuldades enquanto ainda é possível remediar a situação (o que não ocorre nos sistemas de EaD convencionais e nem na maioria das salas de aula presenciais).

Um outro resultado interessante é apresentado por Pimentel e Omar [35]. Neste trabalho, os autores utilizam técnicas da EDM para identificar as relações entre medidas de conhecimento (cognitivas) e medidas metacognitivas. As medidas cognitivas retratam o real desempenho do aluno na resolução de cada problema enquanto que as medidas cognitivas indicam o grau de consciência (*awareness*) do aluno em relação ao seu próprio conhecimento.

A EaD no Brasil também oferece grandes oportunidades para se realizar pesquisas relacionadas ao suporte ao diálogo e a discussão. Conforme resultados obtidos por Scheuer e McLaren [40] é possível utilizar técnicas da área de EDM para apoiar professores a conduzir discussões em salas de aula virtuais de forma efetiva. A EDM também pode proporcionar benefícios à avaliação e a aprendizagem através de discussões assíncronas [cf. 21], utilizando ferramentas que são freqüentemente encontradas nos ambientes de EaD utilizados no Brasil. Uma revisão do estado da arte sobre o potencial da EDM em melhorar os cursos via internet nas universidade é apresentado por Romero, Ventura, & Garcia [38]. Neste trabalho os autores recomendam à todos os pesquisadores interessados que apliquem os métodos disponíveis na área de mineração de dados educacionais para casos específicos de EaD e, dessa forma, promover um ensino mais personalizado e de melhor qualidade.

Um passo importante e, necessário, para que a área de EDM tenha resultados tão positivos no Brasil quanto aqueles obtidos no exterior, será a padronização dos dados obtidos nos cursos de EaD. Estes dados precisam ser sistêmicos, anônimos, e seguindo um padrão bem definido que seja utilizado por todos os ambientes virtuais. É importante que todas as informações necessárias sejam coletadas e que os dados sejam estruturados de forma a considerar os resultados das análises anteriores, pois dessa forma as informações mais relevantes são enfatizadas, melhor compreendidas e mais facilmente utilizadas [cf. 1, 27, 34, 38].

A quantidade de alunos em cursos de EaD cria oportunidades excelentes para pesquisas na área de EDM e pode, futuramente, beneficiar significativamente o processo de ensino e aprendizagem no Brasil. Contudo, o desenvolvimento de pesquisas nesta área vai depender da avaliação de dados pela comunidade científica brasileira, assim como aconteceu com a comunidade internacional que criou o Pittsburgh Science of Learning Center e o

DataShop, como discutido anteriormente neste artigo. No Brasil, acredita-se fortemente que um esforço conjunto envolvendo pesquisadores, educadores e reguladores deva ser realizado para que o progresso nessa área ocorra de forma ágil. Através da estruturação e do armazenamento de dados de alta qualidade será possível disponibilizar publicamente para toda a comunidade de pesquisa brasileira e internacional, grandes quantidades de dados que, se analisadas corretamente, poderão beneficiar estudantes do Brasil e do mundo através de: (a) mecanismos e ferramentas educacionais mais eficiente; (b), modelos para identificar alunos com dificuldades de aprendizagem; (c) meios de melhorar a qualidade do material didático; e (d) o desenvolvimento de métodos pedagógicos mais eficazes; além de outros.

6. Conclusões

A mineração de dados educacionais (EDM) surgiu como uma área de pesquisa que possui grande potencial para contribuir com a melhor compreensão dos processos de ensino, de aprendizagem e de motivação dos alunos tanto em ambientes individuais quanto em ambientes colaborativos de ensino. No momento, as principais contribuições da EDM estão focadas em duas linhas principais: (a) a análise de dados e a criação de modelos para melhor compreender os processos de aprendizagem; e (b) o desenvolvimento de métodos mais eficazes para dar suporte à aprendizagem quando o aluno estuda utilizando softwares educacionais (e.g. cursos via internet). Nos EUA e na Europa diversos sistemas tutores inteligentes estão utilizando técnicas da EDM para proporcionar uma aprendizagem mais personalizada e de melhor qualidade. Ao mesmo tempo, resultados da área já vêm influenciando outros domínios como, por exemplo o ensino de leitura para crianças como apresentado por Beck e Mostow [11]. O Brasil tem uma grande oportunidade para promover a revolução da EDM e beneficiar milhares de alunos; em grande parte por causa do incentivo governamental ao uso da Educação-a-Distância (EaD). Através da coleta de dados em grande escala é possível criar modelos e fazer predições que serão aplicáveis uma qualquer ambiente virtual de aprendizagem e até mesmo em salas de aula convencionais. Para isso, é preciso que os dados das interações dos alunos com o material didático e com os professores e colegas nos ambientes de EaD sejam disponibilizados de forma padronizada e estruturada para comunidade científica brasileira. Além disso, esses dados precisam incluir as informações necessárias para viabilizar a pesquisa e o estudo aprofundado da educação no país. Assim, acredita-se que a EDM tem grande potencial para ajudar o Brasil a se destacar no cenário educacional mundial através de ações que promovam o ensino eficaz nos ambientes de EaD e nas escolas através do uso de tecnologias educacionais que complementam o ensino em sala de aula..

Agradecimentos

Os autores gostariam de agradecer o apoio do Centro de Ciências da Aprendizagem de Pittsburgh (*Pittsburgh Science of Learning Center*) e do apoio da *National Science Foundation* entitulado "*Toward a Decade of PSLC Research*", número de projeto SBE-0836012.

Referências

[1] Allevato, A., Thornton, M., Edwards S., Perez-Quinones, M. Mining data from an automated grading and testing system by adding rich reporting capabilities. In *Proceedings of the International Conference on Educational Data Mining*. 167–176. 2008.

[2] Amershi, S., Conati, C. Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining,* 1(1):18-71. 2009.

[3] Anjewierden, A., Kollöffel, B.J., & Hulshof, C. Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. In *Proceedings of the International Workshop on Applying Data Mining in e-Learning*, pagínas 27-36. 2007.

[4] Baker, R.S.J.d. Data Mining for Education. McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition)*. Oxford, UK: Elsevier. 2010.

[5] Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. Adapting to When Students Game an Intelligent Tutoring System. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. pagínas 392-401. 2006.

[6] Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18 (3): 287-314. 2008.

[7] Baker, R.S.J.d., de Carvalho, A. M. J. A. Labeling Student Behavior Faster and More Precisely with Text Replays. In *Proceedings of the International Conference on Educational Data Mining*. pagínas 38-47. 2008.

[8] Baker, R.S.J.d., de Carvalho, A.M.J.A., Raspat, J., Aleven, V., Corbett, A.T., Koedinger, K.R. Educational Software Features that Encourage and Discourage "Gaming the System". In *Proceedings of the International Conference on Artificial Intelligence in Education*, pagínas 475-482. 2009.

[9] Baker, R.S.J.d., Yacef, K. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1 (1): 3-17. 2009.

[10] Barnes, T., Bitzer, D., Vouk, M. Experimental Analysis of the Q-Matrix Method in Knowledge Discovery. Lecture Notes in Computer Science 3488: Foundations of Intelligent Systems. páginas 603-611. 2005.

[11] Beck, J.E., Mostow, J. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In Proceedings of the International Conference on Intelligent Tutoring Systems. pagínas 353-362. 2008.

[12] Blanchard, E. G., Roy, M., Lajoie, S. P., Frasson, C. An evaluation of sociocultural data for predicting attitudinal tendencies, In Proceedings of the International Conference on Artificial Intelligence in Education, pagínas 399-406, 2009.

[13] Brandão, M. F. R., Ramos, C. R. S., Tróccoli, B. T. Análise de agrupamento de escolas e Núcleos de Tecnologia Educacional: mineração na base de dados de avaliação do Programa Nacional de Informática na Educação, 366-374, 2006.

[14] Cen, H., Koedinger, K., Junker, B. Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. In Proceedings of the International Conference on Intelligent Tutoring Systems. 12-19. 2006.

[15] Cen, H., Koedinger, K., Junker, B. Is over practice necessary? Improving learning efficiency with the cognitive tutor through educational data mining. In Proceedings of the International Conference on Artificial Intelligence in Education. páginas 511-518. 2007.

[16] Corbett, A.T., & Anderson, J.R. Knowledge Tracing: Modeling the Acquisiton of Procedural Knowledge. User Modeling and User-Adapted Interaction, 4: 253-278. 1995.

[17] Desmarais, M.C., Maluf, A., Liu, J. User-expertise modeling with empirically derived probabilistic implication networks. User Modeling and User-Adapted Interaction, 5(3-4): 283-315. 1996.

[18] Desmarais, M.C., Pu, X. A Bayesian Student Model without Hidden Nodes and Its Comparison with Item Response Theory. International Journal of Artificial Intelligence in Education 15: 291-323, 2005.

[19] Dias, M. M., Filho, L. A. S., Lino, A. D. P., Favero, E. L., Ramos, E. M. L. S. Aplicação de Técnicas de Mineração de Dados no Processo de Aprendizagem na Educação a Distância. Anais do Simpósio Brasilerio de Informática na Educação, 105-114, 2008.

[20] Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A. & Fischer, F. Supporting CSCL with automatic corpus analysis technology, In CSCL '05: Proceedings of the 2005 Conference on Computer Support for Collaborative Learning. páginas 125-134. 2005.

[21] Dringus, L.P., Ellis, T. Using data mining as a strategy for assessing asynchronous discussion forums, Computers and Education, 45: 141–160. 2005.

[22] Hershkovitz, A., Nachmias, R. Developing a Log-Based Motivation Measuring Tool. In Proceedings of the International Conference on Educational Data Mining, 226-233. 2008.

[23] Isotani, S., Inaba, A., Ikeda, M. and Mizoguchi, R. (2009) An Ontology Engineering Approach to the Realization of Theory-Driven Group Formation. International Journal of Computer-Supported Collaborative Learning, Springer, 4(4), 445-478.

[24] Kampff, A. J. C. Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. Tese de doutorado em Informática na Educação. Universidade Federal do Rio Grande do Sul, 2009.

[25] Kay, J., Maisonneuve, N., Yacef, K. Reimann, P. The Big Five and Visualisations of Team Work Activity. In Proceedings of Intelligent Tutoring Systems (ITS06). 197-206. 2006.

[26] Klemann, M., Reategui, E., Lorenzatti, A. O Emprego da Ferramenta de Mineração de Textos

SOBEK como Apoio à Produção Textual. Anais do Simpósio Brasilerio de Informática na Educação, 2009.

[27] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. A Data Repository for the EDM commuity: The PSLC DataShop. Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press. 2010.

[28] Longhi , M. T., Behar, P. A., Bercht, M., Simonato, G. Investigando a subjetividade afetiva na comunicação assíncrona de ambientes virtuais de aprendizagem. Anais do Simpósio Brasilerio de Informática na Educação, 2009.

[29] Marinho, T., Dermeval, D., Ferreira, R., Braz, L. M., Bittencourt, I. I., Costa, E. B., Luna, H. P. L. Um Framework para Mineração de Dados Educacionais Basedo em Serviçõs Semânticos. Anais do Simpósio Brasilerio de Informática na Educação, 2009.

[30] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006). 935-940. 2006.

[31] Ministério da Educação do Brasil (MEC). Relatório disponível em: http://portal.mec.gov.br/index.php?option=com_content&view=article&id=13260:novo-mecanismo-de-busca-de-polos-pode-ser-acessado-pela-internet&catid=210. 2009.

[32] Moore, A. Statistical Data Mining Tutorials. Available online at :

http://www.autonlab.org/tutorials/ 2005.

[33] Pavlik, P., Cen, H., Wu, L. and Koedinger, K. Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor. In *Proceedings of the International Conference on Educational Data Mining*, 77-86. 2008.

[34] Prata, D.N., Baker, R.S.J.d., Costa, E., Rosé, C.P., Cui, Y., de Carvalho, A.M.J.B. Detecting and Understanding the Impact of Cognitive and Interpersonal Conflict in Computer Supported Collaborative Learning Environments. In *Proceedings of the International Conference on Educational Data Mining*, 131-140. 2009.

[35] Pimentel, E.P., Omar, N. Descobrindo Conhecimentos em Dados de Avaliacao Aprendizagem com Tecnicas de Mineracao de Dado. Workshop sobre Informática na Escola. *Anais do Congresso da Sociedade Brasileira de Computação*, 147-155, 2006

[36] Preti, O. Educação a Distância: indicios de um percurso. Cuiabá: NEAT/IE – UFMT. 1996.

[37] Romero, C., Ventura, S., Espejo, P.G., Hervas, C. Data Mining Algorithms to Classify Students. In *Proceedings of the International Conference on Educational Data Mining,* 8-17. 2008.

[38] Romero, C., Ventura, S., Garcia, E. Data mining in course management systems: Moodle case study and tutorial, *Computers & Education*, 51: 368–384, 2008.

[39] Scheines, R., Sprites, P., Glymour, C., Meek, C. *Tetrad II: Tools for Discovery.* Lawrence Erlbaum Associates: Hillsdale, NJ. 1994.

[40] Scheuer, O. & McLaren, B.M. Helping teachers handle the flood of data in online student discussions. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (ITS-08). 323-332. 2008.

[41] Witten, I.H., Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, San Francisco, CA.