# Uma Análise de Interação em Fóruns de EAD

**Henrique Pequeno[1], Ricardo L. F. de Ávila[2], Elifranio Cruz[1], Marcondes Alexandre[1], Ernesto Trajano de Lima[1], Miguel Franklin de Castro[1]**

[1]Instituto UFC Virtual – Universidade Federal do Ceará (UFC), [2]Centro Universitário Christus – (Unichristus)

{henrique, elifranio, marcondes, ernesto}@virtual.ufc.br,
ricardo.lims@gmail.com, miguel@ufc.br

***Resumo.*** *O registro de interações de alunos e professores/tutores em Ambientes Virtuais de Aprendizagem oferece importante oportunidade para que pesquisadores utilizem técnicas de estatística e de mineração visando a uma análise inteligente de dados. Dessa forma, professores e gestores acadêmicos podem valer-se dessas análises a fim de auxiliá-los em seus processos de tomada de decisão. O presente trabalho objetiva investigar o padrão de comportamento de alunos e professores de cursos a distância da Universidade Federal do Ceará, quanto às suas interações virtuais nos fóruns de discussão, segundo duas perspectivas de oferta de disciplinas, uma modular (disciplinas sequenciais) e outra semestral (disciplinas simultâneas). Foram utilizadas regras de associação e algoritmos de classificação para analisar o comportamento dos partícipes dos cursos. Os resultados obtidos a partir do uso dessas ferramentas demonstraram a importância de seu uso para amparar gestores acadêmicos em decisões que promovam melhor adequação ao contexto de cada curso.*

***Abstract.*** *The record of students and teachers interactions in Virtual Learning Environments offers important opportunity for researchers using statistical and mining techniques aiming an intelligent data analysis. Therefore, teachers and academic managers can make use of these analyses in order to support their decision processes. The present work aims investigate the pattern of students and teachers behavior in distance educational courses of Federal University of Ceará, concerning their virtual interactions in the discussion forums on two modalities of disciplines offer, including a modular (sequential disciplines) and other semester (simultaneous disciplines). Association rules and classification algorithms were used to analyze the courses participants behavior. The results obtained from the use of these tools have demonstrated the importance of its use to support academic managers in decisions that promote better adaptation to the context of each course.*

## 1. Introdução

A utilização de técnicas estatísticas e computacionais para o processamento e a análise de grandes volumes de dados tem ganhado destaque, sobretudo, pelo uso comercial. São representativos os casos exitosos que fazem uso dessas técnicas para estímulo do desejo de compras das pessoas, ou mesmo para obtenção de insumos para a definição de novas estratégias de negócio [Bramer, 2013]. Vale ressaltar, porém, que diversas outras áreas tem feito uso de análise de dados inteligentes, como educação, saúde e segurança [Elmari e Navathe, 2010]. Notadamente na área educacional, diversos estudos têm sido realizados com o intuito de contribuir direta ou indiretamente com o aprendizado de alunos. Essa área de pesquisa emergente é denominada *Educational Data Mining* (EDM) [Romero e Ventura, 2007; Baker, 2011].

Considerando o crescimento do uso de sistemas de informação em atividades educacionais, como cursos na modalidade de educação a distância (EaD), os registros de dados desses sistemas oferecem à EDM ampla condição de investigação e experimentação

[Gottardo, Kaestner e Noronha, 2012]. Em cursos EaD, por exemplo, é comum a utilização de Ambientes Virtuais de Aprendizagem (AVA) para sua oferta. Para o presente estudo, foi utilizado o AVA SOLAR (Sistema Online de Aprendizagem) desenvolvido pelo Instituto UFC Virtual, da Universidade Federal do Ceará.

Nesses sistemas, é possível utilizar dados, como quantitativos de participação de alunos e professores em ferramentas de comunicação (fórum, *chat*, *webmail*, webconferência), rendimento de discentes, correção de trabalhos etc. Esses dados poderão ser utilizados como subsídios para auxiliar na resposta a diversos questionamentos, como:

- De que forma estão ocorrendo as interações docente-aluno e/ou aluno-aluno? Esta abordagem pode ser aprofundada em [Mazza, 2011].
- Como identificar estudantes com problemas de aprendizagem e participação? Esta indagação é explorada por [Jaggars, 2013].
- Como identificar alunos que estão prestes a evadir o curso ou reprovar por nota/participação? Esta abordagem preditiva é tratada por [Liu e Chen, 2011].
- Qual o padrão de interação em fóruns? Esta questão é discutida por [Jacob e Sam, 2010].
- Como a distribuição de prazos de atividades influencia na participação do aluno?

A última questão listada é objeto de estudo do presente trabalho, o qual consiste na investigação quanto ao impacto do modo de oferta de disciplinas (modular ou semestral) no padrão de interação dos partícipes de cursos de graduação a distância. Essa motivação dá-se em razão de os fóruns de discussão representarem uma ferramenta de extrema importância para boa parte dos cursos de EaD [Azevedo, Behar e Reategui, 2011]. Os fóruns permitem que alunos e professores/tutores se manifestem reflexivamente sobre questões atinentes ao curso, compartilhando experiências e conhecimentos coletivamente, o que nos move a analisar como decisões acadêmico-administrativas sobre o modo de oferta de disciplinas têm auxiliado ou penalizado sua experiência de uso.

A análise dos dados feita por este trabalho utiliza técnicas de estatística (correlação linear) e de mineração de dados (classificadores/J48 e regras de associação/*Apriori*) para analisar a distribuição de mensagens de alunos e professores/tutores dentro do período de aplicação de cada fórum. As técnicas foram utilizadas complementarmente para compor a análise de diferentes elementos observados no contexto investigado.

Este trabalho está estruturado em outras 5 seções. Na seção 2, analisam-se trabalhos relacionados. Na seção 3, apresentam-se os critérios utilizados para seleção e tratamento do conjunto de atributos para representação da participação no Ambiente Virtual. Na seção 4, são apresentados os experimentos realizados. Já na seção 5, os dados coletados são apresentados em tabelas e analisados. Finalmente, na seção 6, apresentam-se as conclusões e as perspectivas de continuidade deste trabalho.

## 2. Trabalhos Relacionados

Uma das técnicas que pode ser utilizada para otimizar os processos de aprendizagem em um AVA é fazer uso do conhecimento obtido a partir de análise de dados históricos, em particular, de informações educacionais nos períodos de interesse. No trabalho de Gottardo, Kaestner e Noronha (2012), por exemplo, foi realizada uma análise de dados sobre as notas obtidas pelos alunos. Em seus experimentos, foram avaliadas as possibilidades de obtenção de inferências sobre o desempenho de estudantes em diferentes etapas de realização do curso, porém não buscava analisar nenhum aspecto referente ao uso de fórum de discussão.

Em Mazza (2011), foi apresentado um trabalho que visa a aprofundar as técnicas de visualização de dados educacionais em ambientes de ensino. Seu objetivo é comprovar que

atividades propostas estão sendo realizadas dentro do prazo estipulado. Entretanto, este trabalho também não aborda aspectos de comportamento em fóruns de discussão.

Em Azevedo, Behar e Reategui (2011), foi desenvolvida uma ferramenta para auxiliar o docente na análise qualitativa das mensagens de fóruns. O *software* permite executar, de forma automatizada, o referido processo. O objetivo dos experimentos deste trabalho foi comparar a média das relevâncias das mensagens, com a média das avaliações das postagens feitas pelos professores. Essa ferramenta se assemelha à técnica de filtragem colaborativa, que prevê a avaliação por parte de usuários para melhoria da experiência de demais membros de um grupo.

Uma pesquisa com amplitude maior foi desenvolvida por Dekker, Pechenizkiy e Vleeshouwers (2009), na qual os autores trabalharam com dados de alunos de graduação do curso presencial de Engenharia Elétrica da Universidade de Eindhoven. Nesse trabalho, foi possível identificar, no primeiro ano letivo, tendências dos alunos com risco de evasão. Os pesquisadores submeteram os dados a diversos algoritmos de classificação existentes na ferramenta de mineração de dados Weka [Hall, Frank, Holmes, Pfahringer, Reutemann e Witten, 2009], alcançando resultados entre 75% e 80% de precisão.

Em Manhães, Cruz, Costa, Zavaleta e Zimbrão (2011), foram avaliadas técnicas de mineração por meio de três experimentos. Neles, foram aplicados dez algoritmos de classificação sobre uma base de dados dos alunos de graduação do curso de Engenharia Civil da UFRJ, sendo observado que o algoritmo baseado em modelos probabilísticos (*Naive Bayes*) obteve melhor acurácia para o experimento em estudo. Esse trabalho não define adequadamente como são escolhidos os critérios e parâmetros de classificação, fato que pode influenciar a ocorrência de vícios probabilísticos na análise das informações dos dados minerados. Outro fato relacionado ao escopo da pesquisa foi que os autores não consideram dados relacionados a fóruns e não envolvem a amplitude de cursos a distância.

## 3. Seleção e Tratamento de Dados

Como apresentado na seção introdutória, o presente artigo tem como objetivo investigar o impacto do modo de oferta de disciplinas (modular ou semestral) no padrão de interação de alunos e professores/tutores nos cursos de graduação a distância. Tal investigação foi realizada com base na análise das postagens e participações em fóruns de discussão de tutores e alunos em cursos de graduação a distância da Universidade Federal do Ceará. Foram utilizados fóruns dos anos de 2011 e 2012, os quais contaram, respectivamente, com ofertas de disciplinas modulares (sequenciais) e semestrais (simultâneas). Para tanto, a modelagem dos dados investigados utilizou uma modelagem dimensional, com agrupamento de seis dimensões e uma tabela fato. A estrutura empregada é descrita a seguir, buscando contemplar diversos aspectos de uso e interação em um AVA:

- **Dimensão tutor**: dados que representam o perfil do tutor de cada disciplina, bem como sua interação com os alunos e participação no ambiente. Foram definidos indicadores gerais de quantidade e tempo médio de acessos aos recursos do AVA, além de atributos que representem atividades rotineiras e regulares dos acessos dos aprendizes.
- **Dimensão graduação**: dados que identificam as atividades acadêmicas do curso em análise, por meio de uma descrição e um código identificador de controle.
- **Dimensão disciplina**: dados úteis para mapear e localizar informações de rotina, bem como sua descrição, período, curso ao qual integra e sua devida identificação. Esta dimensão é útil para inferirmos a associação de comportamentos dos alunos e identificarmos tendências associadas às disciplinas envolvidas.
- **Dimensão aluno**: dados que correspondem aos cadastros de informações acadêmicas sobre os estudantes, tendo aplicação para análise comportamental de identificação, inferência e tratamento adequado, em defesa da aprendizagem de

um grupo, personalizando o saber ensinar.

- **Dimensão polo**: dados que recuperam a informação de onde ocorrem os resultados, representando uma justificativa à estrutura instalada, gerando contexto acadêmico para a localidade de sua instalação.
- **Dimensão tempo**: dados que permeiam as características temporais da ocorrência dos fatos analisados, conduzindo bases fundamentais de conhecimento sobre mudanças e justificativas, análise de impactos positivos e negativos, superação de técnicas legadas e validação de ideias inovadoras, analise da superação operacional ao longo do tempo.
- **Fato participação**: dados que vislumbram a natureza operacional do AVA, bem como o correto funcionamento das atividades, destacando as condutas de alunos e tutores no tocante à participação das atividades do ambiente, de modo específico nas discussões em fóruns. Esta estrutura é fundamental para a geração de contexto acadêmico e a descoberta de padrões de comportamento observados ao longo do fator tempo[1].

A Figura 1 ilustra o processo de análise desenvolvido neste artigo, apresentando, também, as bases de dados utilizadas[2]. Após a carga dos dados e a montagem do cubo de dados, diferentes tipos de relatórios gerenciais, administrativos e/ou acadêmicos podem ser gerados, criando a estrutura de Inteligência Empresarial (ou *Business Intelligence*, em inglês), essenciais para os gestores institucionais.

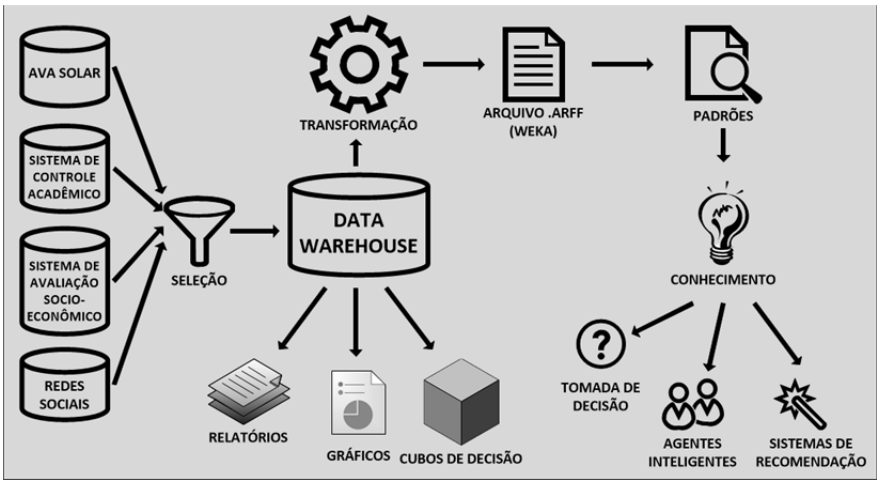


**Figura 1: Modelo de análise utilizado.**

Para os experimentos descritos na Seção 4, durante o processo de transformação ilustrado na Figura 1, foi extraído um conjunto de dados que aferisse a participação de alunos e tutores em fóruns. A escolha dos dados foi feita de forma independente do curso de graduação ou disciplina de tutores e alunos. Inicialmente, foram levantados os 100 fóruns com maior número de participações. Os atributos selecionados foram os seguintes: polo, disciplina, curso, fórum, total de mensagens do fórum, data de postagem, data de início e data de fim do fórum[3].

A partir da composição do cubo de dados, foram realizados alguns testes utilizando algoritmos de regressão linear, regressão não linear e de regras de associação. Os

[1] O banco de dados foi construído por meio de rotinas de extração e transformação de dados do AVA, utilizando o *software* Pentaho (http://www.petaho.com). Os dados, então, foram inicialmente persistidos em um Sistema Gerenciador de Banco de Dados (SGBD) SQL Server, fonte de dados a partir da qual o *Data Warehouse* supra descrito pudesse ser construído no SGBD MySQL.

[2] As informações de redes sociais não foram utilizadas neste trabalho, uma vez que não houve tempo hábil para preparar os dados para o presente artigo.

[3] Os dados resultantes desta extração foram organizados em arquivos no formato .arff, utilizado pela ferramenta WEKA (http://www.cs.waikato.ac.nz/ml/weka). A correta geração destes arquivos é essencial para o Processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases*, ou KDD) que foi definido por Fayyad, Piatetsky-Shapiro e Smith (1996) como "processo não-trivial para identificar padrões válidos, novos, potencialmente úteis e compreensíveis nos dados".

resultados iniciais, entretanto, não foram satisfatórios principalmente devido à pequena quantidade de dados utilizada. Refez-se, então, a extração de forma que, em seu resultado, constassem os 1000 fóruns com maior número de participações. Desse modo, foi possível fazer uma análise mais segura dos resultados. Em termos percentuais, 100 representa cerca de 1% do universo de fóruns, enquanto 1000 representam cerca de 10%.

Ainda durante esta fase inicial de extração e tratamento dos dados, verificou-se que as informações referentes às postagens de tutores e alunos deveriam ser adaptadas. Da forma como foram extraídas, isto é, como números absolutos representando a quantidade total de postagens de tutores e alunos[4], não se podia derivar informação relevante. Dessa forma, estes atributos foram tratados de forma particular.

Considerando o relato de professores/tutores de que postagens dos alunos nos fóruns em momentos muito próximos ao término de seu prazo acarretava prejuízo ao planejamento didático-pedagógico sobre o uso da atividade discursiva, foi realizado um agrupamento dos dados de acordo com os períodos de postagem. Assim, foram definidos três intervalos de tempo para cada fórum.

Na Figura 2, são demonstradas as subdivisões temporais realizadas em cada fórum. O prazo operacional do fórum foi dividido em três períodos de tempo, em que cada intervalo corresponde a um terço do tempo estipulado para realização de um fórum. Foram estabelecidos de forma conceitual, os seguintes critérios limites no decorrer do tempo:

- 1ΔTx - Início
- 2ΔTx – Interm,ediário
- 3ΔTx - Final

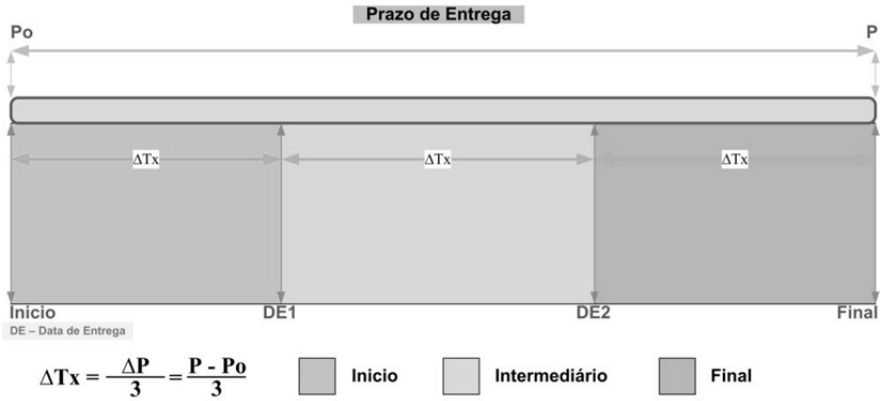


**Figura 2. Divisão proposta para os fóruns**

A partir dessa lógica de agrupamento, os 2.000 principais fóruns (com maior número de postagens) foram organizados, de forma a possibilitar que fosse analisado o padrão de interação de alunos e professores/tutores de acordo com a natureza de oferta da disciplina em que cada fórum estava vinculado.

## 4. Experimentos realizados

Foram realizados dois experimentos com os dados extraídos do *data warehouse*: no primeiro deles, tentou-se inferir regras de associação entre os dados. Já no segundo, tentou-se criar classificadores que determinavam se o fórum estava relacionado a uma disciplina de oferta modular ou a uma de oferta semestral.

Em ambos os experimentos, os dados foram transformados para o formato .arff, utilizado pela ferramenta Weka [Weka, 2014].

[4] Por exemplo, no fórum 1, o tutor postou 20 vezes e os alunos 46 vezes. Já no fórum 12, o tutor postou 50 vezes, enquanto os alunos postaram 12 vezes.

## 4.1. Experimento 1

O primeiro experimento teve como objetivo a identificação de regras de associação entre as instâncias contidas no arquivo .arff., ou seja, verificar se existem associações de causa-efeito entre as instâncias. Desse modo, pode-se gerar argumentos que fundamentam respostas para as questões supracitadas em na primeira seção do artigo. Para tanto, o algoritmo escolhido foi o *Apriori* [Agrawal e Srikant, 1994] devido a sua simplicidade e a rapidez de respostas de associação envolvendo os dados.

Para minerar comportamentos padrões de postagem dos tutores e alunos, foram utilizados os arquivos .arff[5] ilustrados na Figura 3. Nele, os atributos utilizados foram os seguintes: o total de mensagens por período (períodos 1, 2 e 3), o total de mensagens dos tutores em cada período (idem) e o total de mensagens dos alunos em cada período (idem). Foram avaliadas a estrutura modular (2011) e semestral (2012).

```
@relation 2011_total_periodo_top_1000_Aluno

@attribute total_per_01 {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_01_tutor {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_01_aluno {SP, 1-20, 21-60, 61-100, 100>}
@attribute total_per_02 {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_02_tutor {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_02_aluno {SP, 1-20, 21-60, 61-100, 100>}
@attribute total_per_03 {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_03_tutor {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_03_aluno {SP, 1-20, 21-60, 61-100, 100>}

@data

21-60,SP,21-60,21-60,SP,21-60,21-60,1-20,21-60
21-60,1-20,21-60,21-60,1-20,21-60,61-100,1-20,61-100
```

```
@relation 2012_total_periodo_top_1000_Aluno

@attribute total_per_01 {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_01_tutor {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_01_aluno {SP, 1-20, 21-60, 61-100, 100>}
@attribute total_per_02 {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_02_tutor {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_02_aluno {SP, 1-20, 21-60, 61-100, 100>}
@attribute total_per_03 {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_03_tutor {SP, 1-20, 21-60, 61-100, 100>}
@attribute per_03_aluno {SP, 1-20, 21-60, 61-100, 100>}

@data

1-20,SP,1-20,21-60,SP,21-60,100>,1-20,100>
1-20,SP,1-20,1-20,1-20,1-20,100>,1-20,100>
```

**Figura 3. Início dos arquivos .arff utilizados no primeiro experimento**

Os atributos acima descritos eram originalmente numéricos. Para facilitar a geração e o entendimento das regras de associação, entretanto, foi realizada uma discretização dos valores desses atributos. Foram geradas cinco faixas de valores, sendo elas as seguintes: dois *outliers*, um com o valores "SP" (abreviação de "*sem postagem*"), correspondendo ao *outlier* negativo, e outro com o valor "100>", correspondendo ao *outlier* positivo, i.e., com uma grande quantidade de postagens. Os demais valores foram divididos em três faixas: uma com quantidades correspondentes a 1-20 postagens (baixa participação), a 21-60 (média participação) e a 61-100 (boa participação). Salienta-se, no entanto, que, neste trabalho, não foram analisados critérios de qualificação das postagens, ficando esta análise como um trabalho futuro em desenvolvimento. Submetendo o arquivo .arff da Figura 3 ao algoritmo *Apriori*, foi obtido o resultado da Figura 4.

**Regras de associação 2011**
**(Modular)**

1. total_per_01=SP 222 ==> per_01_tutor=SP 222 conf:(1)
2. total_per_01=SP 222 ==> per_01_aluno=SP 222 conf:(1)
3. per_01_tutor=SP per_01_aluno=SP 222 ==> total_per_01=SP 222 conf:(1)
4. total_per_01=SP per_01_aluno=SP 222 ==> per_01_tutor=SP 222 conf:(1)
5. total_per_01=SP per_01_tutor=SP 222 ==> per_01_aluno=SP 222 conf:(1)
6. total_per_01=SP 222 ==> per_01_tutor=SP per_01_aluno=SP 222 conf:(1)
7. total_per_02=1-20 279 ==> per_02_aluno=1-20 273 conf:(0.98)
8. per_02_tutor=1-20 total_per_03=21-60 240 ==> per_03_aluno=21-60 234 conf:(0.98)
9. per_01_tutor=1-20 total_per_03=21-60 222 ==> per_03_aluno=21-60 216 conf:(0.97)
10. total_per_03=21-60 per_03_tutor=1-20 278 ==> per_03_aluno=21-60 266 conf:(0.96)

**Regras de associação 2012**
**(Semestral)**

1. total_per_02=1-20 229 ==> per_02_aluno=1-20 227 conf:(0.99)
2. total_per_01=1-20 per_02_aluno=21-60 219 ==> per_01_aluno=1-20 213 conf:(0.97)
3. total_per_01=1-20 total_per_02=21-60 244 ==> per_01_aluno=1-20 236 conf:(0.97)
4. per_01_tutor=1-20 per_01_aluno=21-60 214 ==> total_per_01=21-60 204 conf:(0.95)
5. total_per_01=1-20 per_02_tutor=1-20 299 ==> per_01_aluno=1-20 284 conf:(0.95)
6. total_per_01=1-20 460 ==> per_01_aluno=1-20 432 conf:(0.94)
7. total_per_01=1-20 per_01_tutor=1-20 per_02_tutor=1-20 228 ==> per_01_aluno=1-20 213 conf:(0.93)
8. total_per_01=1-20 per_03_tutor=1-20 286 ==> per_01_aluno=1-20 267 conf:(0.93)
9. per_01_tutor=1-20 total_per_02=21-60 per_03_tutor=1-20 239 ==> per_02_tutor=1-20 220 conf:(0.92)
10. total_per_01=1-20 per_01_tutor=1-20 335 ==> per_01_aluno=1-20 307 conf:(0.92)

**Figura 4. Resultado gerado com Processamento do arquivo .arff.**

Na Figura 4, fazendo uma análise dos resultados, observa-se, por exemplo, que, no período 1 (regras 3 - modular; regras 7 - semestral), quando os alunos postam pouco (1 a 20 mensagens), nota-se comportamento similar por parte do tutor. Ainda na Figura 4, são exibidas as 10 melhores regras, ordenadas por *cofidence* (sendo o percentual de confiança - *conf*, variando de 0 a 1 como limiares percentuais). Os valores destacados em cinza claro correspondem ao total de ocorrências em que a regra aparece nos treinos estatísticos.

[5] Um arquivo .arff é configurado com o nome da relação (@relation), com a descrição dos atributos (@attribute ) e com os dados (@data), chamados instâncias.

Observa-se que a regra 3 (modular) teve uma ocorrência de 222 e obteve uma confiança de 100% para os registros avaliados. Contudo, a regra 7 (semestral) teve uma predominância de 228 e uma confiança de 93%. Esses resultados preliminares mostram que uma tendência de comportamento por parte dos alunos sofre influência pela capacidade quantitativa de participação por parte do tutor.

## 4.2. Experimento 2

No segundo experimento, considerando que uma instância, com seus atributos, caracteriza um fórum, teve-se como objetivo verificar se seria possível criar um classificador capaz de determinar se uma instância é representação de um fórum de disciplina semestral ou de disciplina modular.

Para este experimento, os algoritmos utilizados foram os seguintes: *Naive Bayes*, *Decision Table*, *OneR* e *J48*. Neste experimento todos os dados numéricos foram normalizados. Tais dados representam o total de dias de um fórum; os totais de mensagens dos períodos 1, 2 e 3; os totais de mensagens dos tutores nos períodos 1, 2 e 3; e os totais de mensagens dos alunos nos períodos 1, 2 e 3. O início do arquivo .arff utilizado encontra-se apresentado na Figura 5.

```
@relation 2011_2012_total_periodos_top_1000_foruns_nominal_com_area_filtrado_com_classe_normalizado

@attribute total_dias_forum real
@attribute total_per_01 real
@attribute per_01_tutor real
@attribute per_01_aluno real
@attribute total_per_02 real
@attribute per_02_tutor real
@attribute per_02_aluno real
@attribute total_per_03 real
@attribute per_03_tutor real
@attribute per_03_aluno real
@attribute total_mensagens_forum real
@attribute graduacao {BAGP,GAD,LQUIM,LLING,LLPT,LMAT,LLESP,LPED,LFIS}
@attribute area {CSA,CET,LLA,CH}
@attribute class {MODULAR,SEMESTRAL}

@data

0.089041,0.159836,0,0.19697,0.147679,0,0.166667,0.134948,0.074766,0.114815,0.188442,BAGP,CSA,MODULAR
0.089041,0.151639,0.045113,0.156566,0.130802,0.022222,0.138095,0.245675,0.074766,0.233333,0.226131,BAGP,CSA, MODULAR
```

**Figura 5. Dados utilizados no experimento 2.**

Os resultados obtidos no experimento encontram-se ilustrados na Tabela 1. Os resultados representam a acurácia, isto é, em termos percentuais, quantas instâncias foram classificadas corretamente. Em todos os casos, durante o treinamento dos classificadores, foi utilizada a validação cruzada (*10-fold*) como método de teste.

**Tabela 1. Resultados obtidos com os classificadores utilizando validação cruzada**

| **Classificador** | *Naive Bayes* | *Decision Table* | *OneR* | *J48* |
|---|---|---|---|---|
| **Acurácia** | 69,3% | 96% | 96,7% | 88,5% |

Para o treinamento realizado, foram utilizados diferentes ajustes no classificador. Foram escolhidas, aleatoriamente, 33% das instâncias para realizar o teste de acurácia. Desse modo, 66% das instâncias são utilizadas para criar o classificador, enquanto 33% são utilizadas para testar a acurácia do classificador. Os resultados dos testes são apresentados na Tabela 2.

**Tabela 2. Resultados obtidos utilizando 33% das instâncias para testar a acurácia**

| **Classificador** | *Naive Bayes* | *Decision Table* | *OneR* | *J48* |
|---|---|---|---|---|
| **Acurácia** | 68,5% | 94,2% | 94,8% | 87,3% |

Nota-se, neste caso, uma ligeira queda de desempenho dos classificadores. Em ambos os casos, entretanto, foi possível criar classificadores com bom desempenho, o que pode indicar um comportamento diferente de tutores e alunos, a depender da forma de oferta de curso (modular ou semestral).

## 5. Análise dos Resultados

Os experimentos realizados mostraram que foi possível identificar regras de associação com altos níveis de confiança (10 regras com níveis de confiança entre 0,91 e 0,97). Foi possível também criar classificadores capazes de diferenciar instâncias que representam fóruns de disciplinas com ofertas modulares e semestrais.

Além do uso desses algoritmos, investigou-se também padrões estatísticos nos dados obtidos a partir do *data warehouse* construído. Por exemplo, observou-se por parte dos alunos uma tendência natural de postagens de trabalhos no intervalo do terceiro período de publicação.

Avaliando as postagens por parte de professores/tutores, frente aos resultados dos alunos, observou-se que os tutores também possuem comportamento semelhante no que tange à distribuição das postagens por período.

Na Figura 6, é mostrada, de forma gráfica, a distribuição de postagens de alunos e professores/tutores nos três períodos estabelecidos para interação nos fóruns.

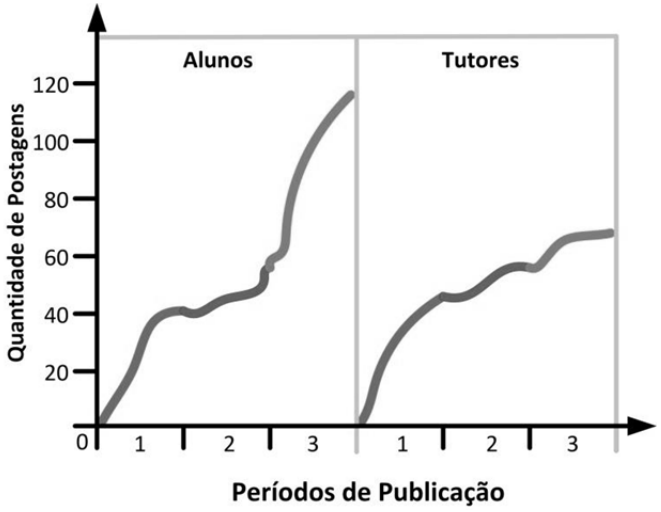


**Figura 6. Gráfico de postagens dos alunos/tutores dentro dos três períodos.**

Vale observar que, considerando os totais de postagens de tutores e alunos, não se encontrou correlação linear (medida pelo coeficiente de Pearson). Nas ofertas modulares, o coeficiente de Pearson foi de 0,23. Já nas ofertas semestrais, o coeficiente calculado foi de 0,22. Isto é, em ambos os casos, uma baixa correlação. Quando se observa, entretanto, os coeficientes por período, a relação se modifica (Tabela 3).

**Tabela 3. Coeficientes de correlação de Pearson por período**

| Período/Oferta | Coeficiente Pearson |
|---|---|
| Per. 1 tutores x alunos (oferta modular) | 0,59 |
| Per. 1 tutores x alunos (oferta semestral) | 0,66 |
| Per. 2 tutores x alunos (oferta modular) | 0,44 |
| Per. 2 tutores x alunos (oferta semestral) | 0,42 |
| Per. 3 tutores x alunos (oferta modular) | 0,26 |
| Per. 3 tutores x alunos (oferta semestral) | 0,16 |

Neste caso, observa-se claramente uma correlação no primeiro período, tanto na oferta modular quanto na oferta semestral. Essa correlação vai diminuindo com o passar do tempo, praticamente inexistindo no terceiro e no último período. Esse fato demonstra que alunos têm deixado para o último momento sua participação nos fóruns. A participação dos tutores não acompanha esse crescimento de postagens dos alunos, pois tanto alunos como tutores não podem mais postar mensagens ao término do prazo previsto para o fórum. Assim, o professor/tutor não tem tempo hábil para estabelecer um *feedback* ao aluno, uma anomalia causada pelas postagens tardias dos alunos.

Interessante ainda notar que, nas ofertas modulares, entre os 1000 fóruns com mais postagens, 48,9% representam fóruns de cursos da área de Linguística, Letras e Artes, enquanto que apenas 22,6% dos fóruns pertencem aos cursos de Ciências Exatas e da Terra. As outras duas áreas, Ciências Sociais Aplicadas e Ciências Humanas, possuem 23,5% e 5%, respectivamente. Já na oferta semestral, a proporção se modifica: 35,7% dos fóruns são de cursos de Linguística, Letras e Artes, enquanto que 42,2% são de cursos da área de Ciências Exatas e da Terra. Nas outras áreas, o percentual é de 19,8% para Ciências Sociais Aplicadas e de 2,3% para as Ciências Humanas.

Esses números refletem a expectativa existente entre alunos e professores de que os alunos deixam para fazer suas atividades nos momentos finais, assim como de que cursos nas áreas de Ciências preferem mais tempo para a realização das atividades (disciplinas semestrais) e os da área de Ciências Humanas preferem atividades mais curtas, porém sequenciais.

## 6. Conclusões e Trabalhos Futuros

Os resultados obtidos neste trabalho demonstram que a utilização de classificadores e regras de associação para verificar a participação de professores e tutores em fóruns de discussão de cursos a distância foi uma exitosa escolha, pois permitiu oferecer análises mais substanciais sobre questões antes não comprovadas, que repousavam no empirismo de opiniões de alunos e professores.

As análises mostraram que para os cursos da área de Linguística, Letras e Artes, é aconselhável que a oferta de suas disciplinas sejam modulares, pois a participação de alunos e professores foi maior quando deste modo de oferta. Já cursos da área de Ciências Exatas e da Terra demonstraram o inverso, ou seja, que o modo de oferta semestral de disciplinas são mais profícuos à participação nos fóruns. Considerando que Fórum consiste em uma das principais ferramentas didático-pedagógicas previstas pelos cursos a distância no ambiente pesquisado, os resultados serão compartilhados com os coordenadores dos cursos para que possam subsidiar suas decisões quanto ao melhor formato de oferta para seus respectivos cursos. Outra importante constatação feita, é que os cursos devem trabalhar fortemente conscientização sobre o uso dos fóruns como um verdadeiro espaço de discussão e não como um repositório para se postar um comentário já muito próximo a seu fechamento. Como visto por este trabalho, esta atitude tem prejudicado a participação dos professores/tutores no estabelecimento de discussões e/ou resposta de questionamentos dos alunos, pois, quando se encerra o prazo do fórum, nem aluno, nem professor podem mais submeter novos comentários. Assim, este trabalho contribui para que os coordenadores de cursos a distância possam tomar decisões a partir das análises realizadas cientificamente.

O trabalho de análise inteligente de dados sobre fontes de dados educacionais tende a se expandir em diversas vertentes e, dessa forma, as instituições de ensino poderão dar saltos de qualidade a partir de ajustes fortemente subsidiados por técnicas de estatística e/ou de mineração de dados.

Como trabalhos futuros, espera-se um aprofundamento das análises sobre as interações dos fóruns, verificando questões também atreladas ao aspecto qualitativo das mensagens, como a investigação das diversas categorias de postagens de tutores e alunos (motivacionais, de conteúdo, referências, reclamações etc.), analisando a significância dessas postagens dos tutores para o bom rendimento do aluno. Para essa análise, será testado o uso de técnicas de pré-processamento de textos em conjunto com algoritmos de busca e comparação textual. [Ávila e Marques, 2013].

## Referências Bibliográficas

Agrawal, R.; Srikant, R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.

Ávila, R.; Soares, J. Uso de técnicas de pré-processamento textual e algoritmos de comparação como suporte à correção de questões dissertativas: experimentos, análises e contribuições. In: Anais do Simpósio Brasileiro de Informática na Educação. Vol. 24. No. 1. 2013.

Azevedo, B. F. T.; Behar, P. A.; Reategui, E. B.; Art. Análise das mensagens de fóruns de discussão através de um software para mineração de textos. 22º Simpósio Brasileiro de Informática na Educação – SBIE - XVII WIE. 2011.

Baker, R. S. J.; Isotani, S.; Carvalho, A. M. J. B. Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, Volume 19, Número 2, 2011.

Bramer, Max. Principles of Data Mining (Undergraduate Topics in Computer Science). Springer; 2nd ed. 2013 edition (February 25, 2013).

Dekker G.; Pechenizkiy M. and Vleeshouwers J. "Predicting Students Drop Out: A Case Study". In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, 2009. Pages 41-50.

Elmari, R.; Navathe, S. B.; Sistemas de Banco de Dados. 6ª Edição. Editora Pearson. (pp 698 – 731). 2010.

Fayyad, U.; Piatetsky-Shapiro, G.; Smith, P. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press/ The MIT Press, MIT, Cambridge, Massachusets, England, 1996.

Gottardo, E.; Kaestner, C.; Noronha, R. V.; Art. Previsão de desempenho de estudantes em cursos EaD utilizando mineração de dados: uma estratégia baseada em séries temporais. 23º Simpósio Brasileiro de Informática na Educação – SBIE. 2012.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. and Witten. I.H. (2003) "The WEKA Data Mining Software: An Update" SIGKDD Explorations, Volume 11. 2009, Issue 1.

Jacob, S. M.; Sam, H. K.; Analysis of interaction patterns and scaffolding practices in online discussion forums; 4th International Conference on Distance Learning and Education (ICDLE), IEEE. 2010.

Jaggars, S. S.; Xu, D.; (2013) Predicting Online Student Outcomes From a Measure of Course Quality. Community college reasearch Center - CCRC Working Paper No. 57. Colegiado de professores – Universidade de Columbia. Abril de 2013.

Liu, K.F.-R.; Chen, Jia-Shen; (2011) Prediction and assessment of student learning outcomes in calculus a decision support of integrating data mining and Bayesian belief networks. IEEExplore. 3rd International Conference on Computer Research and Development (ICCRD). 2011.

Manhães, L. M. B.; Cruz, S. M. S.; Costa, R. J. M.; Zavaleta, J.; Zimbrão G.; (2011) "Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados." 22º Simpósio Brasileiro de Informática na Educação – SBIE - XVII WIE. 2011.

Mazza, R.; Visualization in Educational Environments. Handbook of Educational Data Mining. Ed. CRC Press. 2011.

Romero, C., Ventura, S.; Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135-146. 2007.

Weka. Machine Learning. Disponível em: <http://www.cs.waikato.ac.nz/~ml/weka/index.html>. Acesso em: 07 jan. 2014.