Aplicação de Mineração de Dados no Censo da Educação Superior: Estudo de Caso na Base de Docentes de 2014

Bruno Boavetura de Oliveira Lacerda[1], Manuela Amaral de Araújo[2], Saionara da Silva Araújo[3], Pablo Freire Matos[4]

[1]Discente de graduação em Sistemas de Informação - IFBA. e-mail: brunoboaventura@gmail.com; [2] Discente de graduação em Bacharelado em Sistemas de Informação - IFBA. e-mail: mn21922192 @gmail.com; [3] Discente de graduação em Bacharelado em Sistemas de Informação. e-mail: narabdo@gmail.com; [4] Professor do curso Bacharelado em Sistemas de Informação. e-mail: pablofmatos@ifba.edu.br

RESUMO: Esse artigo tem o objetivo de apresentar uma aplicação da descoberta de conhecimento através da mineração de dados, utilizando o algoritmo Apriori, a ferramenta Weka e soluções/problema envolvendo o tratamento da base de dados. Também são descritos trabalhos correlatos que abordam o tema em questão. Como base de dados, foi selecionada uma que passou por avaliações, pesquisas e exames realizados pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) com relação aos dados do Censo da Educação Superior. Finalmente, é apresentado o resultado das regras de associação obtidas por meio da mineração de dados.
Palavras–chave: análise de dados, descoberta de informação, regras de associação

Data Mining on Higher Education Census: Case Study on 2014 Teacher's Database

ABSTRACT: This paper aims at presenting an application of knowledge discovery through data mining, using the Apriori algorithm, the Weka tool and solutions/problem involving the database's treatment. It also describes related works on this topic. It was selected a database that has passed by ratings, research and tests, carried out by INEP (National Institute of Educational Studies Anísio Teixeira) with respect to the Higher Education Census data. Finally, it shows the result of the association rules obtained using the data mining process.
KEYWORDS: data analysis, information discovery, association rules

INTRODUÇÃO

Com o avanço e redução de custos da tecnologia, é cada vez mais rápido o aumento da quantidade de dados gerados. Alguns desses dados poderiam, se bem analisados, agregar informações importantes para empresas, governos e pessoas. A mineração de dados é uma tarefa muito utilizada para essa finalidade, pois, como parte de um processo de descoberta de conhecimento (KDD), busca gerar informações baseadas em dados gerados por determinado "meio" (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Alguns trabalhos com objetivos semelhantes ao proposto neste artigo podem ser observados na literatura. Fonseca e Namen (2016) utilizaram a técnica de classificação,

através do algoritmo Naïve Bayes, para relacionar características de professores de matemática com o aprendizado dos alunos na disciplina e identificou fatores positivos e negativos em relação ao aprendizado dos alunos. Alguns dos pontos positivos foram o alto percentual de cumprimento do planejamento das aulas, baixo índice de faltas dos professores e assiduidade dos alunos. Já como pontos negativos pode-se citar a desvalorização salarial do professor, a grande quantidade de falta dos alunos.

Outro trabalho nesse campo buscou apresentar uma análise sobre alunos que ingressavam ou saiam das Instituições de Ensino Superior, aplicando técnicas de associação, classificação e agrupamento (PASTA, 2011). Como resultados obtidos, pode-se observar que a maioria dos ingressantes veio de escolas públicas, e escolheu a instituição por conta de sua localização, além de pretender abrir seu próprio negócio. Em relação ao que sai da IES, a escolha da instituição foi influenciada pela matriz curricular dos cursos pretendidos. Um ponto importante a ser observado pela gestão da instituição é que a maioria dos egressos não indicaria a instituição.

Um estudo sobre o aprendizado de língua portuguesa foi o foco de Namen e Soares (2011). Nesse trabalho, os autores trouxeram associações obtidas na aplicação do algoritmo Apriori sobre uma base de dados de uma escola de ensino fundamental do Rio de Janeiro. Com os resultados da pesquisa observou-se que a falta de incentivo dos pais, ou caso o aluno exerça alguma atividade de trabalho são fatores que, dentre outros, afetam o aprendizado do estudante.

Nesse sentido, este trabalho busca demonstrar os resultados obtidos a partir da análise da base de dados do Censo da Educação Superior disponibilizada pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) com aplicação de algoritmo de regras de associação e com a utilização da ferramenta Weka (2016).

Com a aplicação de tarefas de mineração de dados, padrões são descobertos de forma mais automática, mais rápida e com maior grau de certeza. A escolha do método de mineração utilizado é fundamental para a geração de conhecimento. Dentre esses métodos, pode-se citar alguns: Classificação, Agrupamento, Associação e Regressão (CAMARGO *et al.*, 2016).

Para o trabalho proposto, a técnica escolhida foi a de associação, indicada quando se deseja descobrir regras que indiquem as relações entre os atributos dos dados informados (SILVA, 2005). As associações descobertas, também chamadas de regras, são no formato X

⇒ Y, indicando que quando ocorre X também ocorre Y, sendo o X o determinante da regra e Y o resultante (SILVA, 2004). Para gerar essas regras, as medidas usadas são, basicamente, o suporte e a confiança. O primeiro consiste no cálculo de quantas transações onde ocorre X e Y estão no conjunto completo de dados, enquanto o segundo é dado pelo número de ocorrências de X e Y sempre que ocorre X (SILVA, 2005). Dentre os algoritmos de mineração de dados, o utilizado neste trabalho foi o Apriori, que identifica regras de associação entre atributos da base de dados (SILVA, 2005).

Várias são as ferramentas desenvolvidas para tarefas de mineração. A ferramenta Weka foi a escolhida para a realização desse trabalho. Weka é uma ferramenta para aprendizagem de máquina e permite que o computador analise grande quantidade de dados e decida, por meio de algoritmos, quais as informações mais relevantes.

MATERIAL E MÉTODOS

Proposta do trabalho

Este trabalho tem o objetivo de, através do conceito sobre o processo de descoberta de conhecimento, aplicar a teoria no desenvolvimento de um projeto prático utilizando a ferramenta Weka e a Associação como técnica de mineração de dados. Para isso, foi preciso realizar as seguir etapas:

1. Realizar uma busca por uma base de dados recente (a partir de 2010) com uma quantidade razoável de dados;
2. Realizar seleção e transformação de atributos para obter regras de associação relevantes;
3. Utilizar a ferramenta Weka para analisar a base de dados.

Para escolher uma base de dados, diversas consultas foram feitas sem sucesso, pois a maior parte das bases de dados são proprietárias e poucas são disponibilizadas ao público em geral, devido às questões de segurança e sigilo. Após conseguir bases no site do INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira[1], foram necessários vários testes, seleção de atributos e transformação em diversos bancos de dados até chegar ao

[1] INEP: http://portal.inep.gov.br

objetivo proposto pelo trabalho que é: selecionar uma base de dados a fim de encontrar informações relevantes não descobertas outrora. Esse processo foi o mais custoso e trabalhoso, pois as bases encontradas muitas vezes retornavam associações óbvias e que não seriam relevantes para nenhuma tomada de decisão.

Processamento e Análise dos Dados

O processamento e análise dos dados foram realizados utilizando as etapas do processo de KDD proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996).

Seleção

A base de dados escolhida foi a do Censo da Educação Superior[2], realizado pelo INEP, disponível em 2014. Como o censo se trata de um arquivo composto de várias bases, foi escolhida a que lista os docentes do ensino superior em todo o Brasil. A mais atual encontrada foi do ano de 2014 e possui 39.659 instâncias com 50 atributos.

Pré-processamento

São necessárias algumas configurações na interface do Weka para que os arquivos possam ser "lidos". Um dos erros frequentes ao tentar abrir um arquivo é a codificação de caracteres. Quando o arquivo possui acentos e caracteres especiais é preciso realizar a conversão para a codificação UTF-8. No caso deste trabalho, foi utilizada a ferramenta *iconv*, disponível na grande maioria dos sistemas UNIX, e presente em todas as distribuições do sistema Linux, sendo o seu uso discriminado a seguir: na linha de comando "iconv -f ISO_8859-1 -t utf-8 DM_DOCENTE.CSV -o docentes-superior-utf8.csv" (sem aspas), onde o "DM_DOCENTE.CSV" foi o arquivo da base de dados antes da formatação e o "docentes-superior-utf8.csv" o nome do arquivo após a conversão.

Um dos requisitos para executar o algoritmo *Apriori* no Weka é não ter atributos numéricos. Logo, foi necessária a realização de uma categorização de vários dos atributos que continham valores 0 e 1 para "não" e "sim", respectivamente. Para essa tarefa foi usado um

[2] Censo da Educação Superior: http://download.inep.gov.br/microdados/microdados_censo_superior_2014.zip

script na linguagem de programação Python (Figura 1) que também foi aproveitado para remover atributos que representavam simplesmente uma discretização de outros.

O *script* trata os dados seguindo a seguinte sequência: para cada registro, imprime na tela a instância já tratada, com os atributos não selecionados já removidos e categoriza os atributos conforme a necessidade. Os valores dos atributos numéricos 0 e 1 são substituídos por N e S, respectivamente (Linhas 9 a 12, Figura 1). Os atributos categóricos não são alterados. Através do operador de redirecionamento ">" (sem as aspas), a saída do *script* é escrita em um novo arquivo que conterá os dados já tratados.

```
1  # uso do script:
2  # python3 tratar-dados.py > NOME_DO_NOVO_ARQUIVO
3
4  import csv
5
6  def categorizar (valor):
7      try:
8          teste = int (valor)
9          if (teste == 0):
10             texto = "N"
11         else:
12             texto = "S"
13     except:
14         texto = valor
15
16     return texto
17
18 if __name__ == '__main__':
19     # abre o arquivo CSV
20     with open('docentes-superior-utf8.csv', 'r') as f:
21         # cria o objeto que vai ler os registros com separador "|"
22         reader = csv.reader(f, delimiter='|')
23         # para cada linha, gera somente novos registros com os
24         # atributos selecionados, com separador ";"
25         for linha in reader:
26
27             # Campos usados (base zero - lista iniciada em 0)
28             for i in [1,3,5,6,10,12,14,16,22,25,37,38,39,40,41,42,43,44,45,46,47,48]:
29                 valor = categorizar(linha[i])
30                 # Usa ? em valores nulos
31                 print ('?' if (valor =='') else valor, end=';')
32             print ('')
```

Figura 1. *Script* Python (comentado) utilizado para tratamento dos dados. IFBA, 2016.

Formatação

Ao tentar abrir a base de dados no Weka, podem ocorrer algumas mensagens de erro, devido às configurações conflitantes com o padrão de geração do arquivo, já que a ferramenta prevê a vírgula (,) como separador de atributo e o script gerou arquivos com ponto e vírgula (;) como separador. Para isso, antes de clicar em abrir, é preciso visualizar a opção "*invoke options dialog*", opção que vem, por padrão, desmarcada.

Caso não seja selecionada e o arquivo possuir erros de sintaxe, o programa não prossegue. Após clicar em abrir, será disponibilizada uma tela onde a opção

"*enclosureCharacters*" deve ser preenchida somente com "," (sem as aspas) e o campo "*fieldSeparator*" deve ser preenchido com ";" (sem as aspas).

Mineração de Dados

Para conseguir obter regras de associação válidas, é necessário remover alguns atributos e testar os resultados com valores diferentes para o suporte e a confiança. O algoritmo não fica ativo para alguns atributos numéricos, pois ele só trabalha com dados discretos e nominais (SANTOS, 2005). Portanto, é preciso removê-los. A própria interface do Weka, ao abrir o banco de dados, disponibiliza informações sobre os atributos na página inicial. Dos 50 atributos, foram selecionados 12, conforme mostrados na Tabela1.

Tabela 1. Atributos selecionados no pré-processamento. IFBA, 2016.

Atributo	Significado
NO_IES	Nome da IES
DS_CATEGORIA_ADMINISTRATIVA	Nome da Categoria Administrativa
DS_ORGANIZACAO_ACADEMICA	Nome da Organização Acadêmica
IN_CAPITAL_IES	Informa se a IES (reitoria/sede administrativa) está localizada na capital
DS_SITUACAO_DOCENTE	Nome da situação do docente na IES
DS_ESCOLARIDADE_DOCENTE	Informa o nome do grau de escolaridade do docente
DS_REGIME_TRABALHO	Nome do regime de trabalho do docente
DS_SEXO_DOCENTE	Nome do sexo do docente
DS_COR_RACA_DOCENTE	Nome da cor/raça do docente
IN_ATU_EXTENSAO	Informa se o docente atua em atividades de extensão
IN_ATU_PESQUISA	Informa se o docente atua em pesquisa no âmbito de projetos e programas da IES
IN_BOLSA_PESQUISA	Informa se o docente possui bolsa de pesquisa

O algoritmo Apriori é baseado em suporte e confiança, sendo de suma importância a escolha desses parâmetros para a descoberta de regras de associação relevantes. No Weka, a configuração do Apriori: foi realizada da seguinte maneira: lowerBoundMinSupport = 0.40, metricType = Confidence, minMetric = 0.99 e numRules = 30.

RESULTADOS E DISCUSSÃO

Através das regras obtidas (Figura 2), é possível demonstrar o quanto os professores do ensino superior em exercício pouco exercem atividades de pesquisa e extensão. Essa informação pode ser utilizada para que alguma medida seja realizada com o objetivo de incentivar e alterar essa realidade presente no ensino superior.

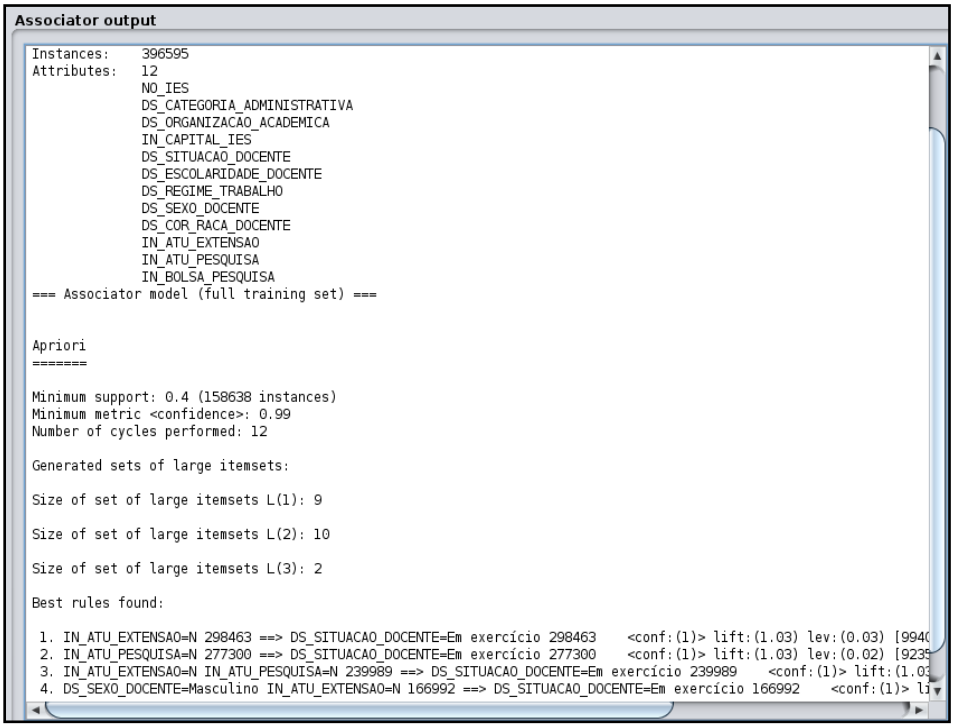

Figura 2. Tela do Weka com as regras encontradas. IFBA, 2016.

A regra IN_ATU_EXTENSAO=N ⇒ DS_SITUACAO_DOCENTE=Em exercício mostra que 298.463 (75,25% do total) docentes em exercício não atuam em projetos de extensão. E de acordo com a regra IN_ATU_PESQUISA=N ⇒ DS_SITUACAO_DOCENTE=Em exercício, a atuação dos docentes em pesquisa é de apenas 30,07%. A terceira regra, IN_ATU_EXTENSAO=N IN_ATU_PESQUISA=N ⇒ DS_SITUACAO_DOCENTE=Em exercício deixa evidente que 60,50% dos docentes não atuam nem em pesquisa e nem em extensão.

CONCLUSÕES

A mineração de dados é um valioso recurso para obtenção de conhecimento. Principalmente na atualidade, onde a quantidade de dados aumenta de forma exponencial e a utilização somente de recursos humanos para análise desses dados se torna insuficiente. As ferramentas desenvolvidas, como o Weka, disponibilizam um recurso importante para análise e obtenção de informações valiosas para a tomada de decisão em diversos ramos empresariais e acadêmicos.

Futuramente, se poderia aplicar essa pesquisa em outros dados disponibilizados no próprio site do INEP, para que os dados sejam utilizados não só para consulta e estatística e sim como fonte de conhecimento para identificar comportamentos comuns que implicam em determinados resultados que, se identificados, poderiam ser um diferencial na tomada de decisões na educação básica e superior. É importante ressaltar que outros atributos poderiam ser utilizados, e outras regras relevantes poderiam ser descobertas, o que pode ser feito em trabalhos futuros.

REFERÊNCIAS

CAMARGO, A. *et al.* Mineração de dados eleitorais: descoberta de padrões de candidatos a vereador na região da campanha do Rio Grande do Sul. **Revista Brasileira de Computação Aplicada**, Passo Fundo, v. 8, n. 1, p.64-73, 2016.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. AAAIIntelligence, p. 37-54, 1996.

FONSECA, S. O; NAMEN, A. A. Mineração em Bases de Dados do INEP: Uma Análise Exploratória para Nortear Melhorias no Sistema Educacional Brasileiro. **Educação em Revista**, BH, v. 32, n. 1, p.133-157, 2016.

NAMEN, A. A; SOARES, A.C. S. Mineração de dados relacionados ao aprendizado de língua portuguesa: um estudo exploratório. In: ENCONTRO DE MODELAGEM COMPUTACIONAL, 14., 2011, Nova Friburgo. **Anais...** Nova Friburgo: Rede Sirius, 2001. p. 295 - 304.

PASTA, A. **Aplicação da técnica de data mining na base de dados do ambiente de gestão educacional:** um estudo de caso de uma instituição de ensino superior de blumenau-sc. 2011. 153 f. Dissertação (Mestrado) - Curso de Computação Aplicada, Universidade do Vale do ItajaÍ, São José, 2011.

SANTOS, R. **Weka na munheca**. 2005. Disponível em: <www.ambientelivre.com.br/downloads/doc_download/81-weka-na-munheca.html>. Acesso em 15 jul. 2016.

SILVA, G. C. **Mineração de regras de associação aplicada a dados da secretaria municipal de saúde de Londrina - PR**. 94 f. Dissertação de Mestrado - UFRGS, Porto Alegre, 2005.

SILVA, M. P. S. **Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka**. 2004. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>. Acesso em 15 jul. 2016.

WEKA. **Machine Learning Group at the University of Waikato:** Downloading and installing Weka. 2016. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>. Acesso em 15 jul. 2016.