INFORMAÇÕES GERAIS DO TRABALHO

Título do Trabalho: Extraindo Conhecimento de Dados Educacionais por meio de Técnicas de Mineração de Dados.

Autor (es): Tares Liberato Orlande de Almeida, Cristiane Norbiato Targa, Bruno Nonato.
Palavras-chave: Mineração de dados, Regras de Associação, Ferramentas, Dados Acadêmicos.
Campus: Sabará.
Área do Conhecimento (CNPq): Ciência da Computação.

RESUMO

A Mineração de Dados ou Data Mining é uma das fases do processo de KDD (Knowledge Discovery in Databases) que objetiva extrair informações, sem conhecimento imediato, de uma grande base de dados para tomada de decisões. Esta metodologia está sendo cada vez mais aplicada em diversas áreas que utilizam o conhecimento para auxiliar os gestores em suas decisões, como empresas, indústrias e instituições de pesquisa. Uma importante informação extraída por meio do processo de mineração são as regras de associação, tais regras representam padrões de relacionamento entre itens de uma determinada base de dados. Instituições de ensino vêm automatizando seus processos, adotando sistemas que permitem capturar e gerir dados pertencentes ao meio acadêmico. Ou seja, também geram e armazenam uma quantidade significativa de dados que podem ser melhor utilizados para aperfeiçoar seus processos de tomada de decisões. O Instituto Federal de Minas Gerais - IFMG - é uma instituição de educação superior, básica e profissional, pluricurricular e multicampi, especializada na oferta de educação profissional e tecnológica nas diferentes modalidades de ensino e utiliza um ERP (Enterprise Resource Planning), conhecido como Conecta, que integra todos os dados e processos envolvidos no contexto acadêmico. Entretanto, os dados gerados e armazenados não são utilizados de forma eficiente para auxiliar no planejamento estratégico. Desta forma, o presente trabalho tem como principal objetivo conhecer e prever o desempenho dos estudantes através da extração de Regras de Associação em dados do Curso de Bacharelado de Sistemas de Informação (BSI) do IFMG - campus Sabará. Para alcançar tal objetivo, realizou-se uma investigação de ferramentas de código aberto que se adequam ao contexto e aos dados do problema em questão, a saber: Weka e RapidMiner. O presente projeto foi inicializado em dezembro de 2017, e ainda está em andamento.

INTRODUÇÃO

O sucesso de uma corporação depende da capacidade de tomar decisões que gerem resultados positivos. O processo de tomada de decisões é responsável pela escolha de uma boa solução para um problema, tornando-se, assim, um ponto chave dentro da gestão organizacional. Para tomar decisões mais assertivas recorrem-se aos dados gerados e armazenados ao longo dos anos pela automatização dos processos organizacionais. O fato de apenas coletar e armazenar os dados não traz nenhuma contribuição para a melhoria na tomada de decisão. Deve-se fazer uma análise para descobrir padrões de comportamento escondidos na base de dados. A mineração de dados possui o objetivo de oferecer estratégias para a análise de grandes bases de dados, procurando extrair informações que estejam implícitas e que sejam desconhecidas e úteis. As regras de associação representam um importante tipo de informação que pode ser obtida através de técnicas de mineração de dados. Estas regras representam padrões de relacionamento entre itens de uma determinada base de dados, tendo como um dos principais objetivos dar apoio à tomada de decisões.

Uma regra de associação representa um padrão de relacionamento entre itens de dados do domínio da aplicação que ocorrem com uma determinada frequência na base de dados, podendo também ser definida como uma implicação da forma X -> Y, onde X e Y são conjuntos de itens pertencentes ao domínio da aplicação tais que $X \cap Y = \not\subset$. X é dito o antecedente e Y, o conseqüente da regra. A interseção vazia entre antecedente e conseqüente da regra assegura que não sejam extraídas regras óbvias que indiquem que um item está associado a ele próprio (Goldschmidt; Passos, 2005).

Como forma de contextualização prática, as regras de associação são amplamente utilizadas na análise de transações de compras, market basket analysis (análise de cestas de compras), tendo como exemplo clássico a análise desenvolvida por uma grande rede de supermercados. A empresa descobriu que o perfil de consumidor de cervejas era semelhante ao de fraldas: homens casados, entre 25 e 30 anos, que compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa. Com base na verificação dessas hipóteses, optou-se por uma otimização das atividades junto às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas (Silva, 2015). A regra de associação fralda -> cerveja indica que o cliente que compra fralda, com um determinado grau de certeza, também compra cerveja.

Dependendo da aplicação, uma transação pode representar os produtos comprados por um determinado cliente ou pode representar as doenças apresentadas por um paciente. A quantidade de regras de associação que podem ser encontradas em uma aplicação de associação é extensa e muitas destas regras não são consideradas relevantes. Uma forma de resolver esta questão é a introdução de medidas de interesse, que fazem a distinção entre as regras relevantes e as não relevantes.

Encontram-se atualmente uma grande oferta de ferramentas que extraem regras de associação em um ambiente com código aberto, como, por exemplo, as ferramentas *Rapid Miner, Orange Data Mining, Weka* entre outras. Apesar de existirem diversas ferramentas que auxiliam na extração de regras de associação, os resultados ainda precisam de uma análise humana para a tomada de decisão.

Assim como as demais organizações, instituições de ensino vêm automatizando seus processos, adotando sistemas que permitem capturar e gerir dados pertencentes ao meio acadêmico. Ou seja, também geram e armazenam uma quantidade significativa de dados que podem ser melhor utilizados para otimizar seus processos de tomada de decisões. O Instituto Federal de Minas Gerais - IFMG - campus Sabará utiliza um ERP *(Enterprise Resource Planning*), sistema denominado Conecta, que integra todos os dados e processos envolvidos no contexto acadêmico. Entretanto, os dados gerados e armazenados não são utilizados de forma eficiente para auxiliar no planejamento estratégico para melhoria dos processos. Desta forma, a utilização de técnicas de mineração de dados, como por exemplo a extração de regras de associação, pode ser utilizada sobre a base de dados dos alunos, buscando estreitar o contato e prever o desempenho acadêmico dos mesmos, disponibilizando assim informações úteis que podem auxiliar no processo de tomada de decisões.

METODOLOGIA

O presente projeto utiliza dados de estudantes do curso de Bacharelado em Sistemas de Informação (BSI) do IFMG - Campus Sabará e foi dividido em etapas de acordo com o KDD - *Knowledge Discovery in Databases* (Fayyad, 1996). O KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. Esse processo é composto pelas etapas:

- Seleção de Dados: consiste na escolha de qual conjunto de dados será submetido ao processo. Os dados foram fornecidos no formato ".XLS", de forma que cada linha da tabela contém os dados de um aluno. Então é a partir dessa etapa que os dados são selecionados, de modo que mantenham a confidencialidade das amostras, os campos **Nome**, **Telefone** e **Endereço** foram dispensados.

- Pré-processamento: fase de reparação, organização e tratamento dos dados. Isso quer dizer que os dados são modificados para se tornarem mais fáceis de manipular. Como exemplo pode-se citar o ocorrido com o campo correspondente à **Data de Nascimento**, tal dado transformado em um novo campo **Idade** que foi posteriormente discretizado em intervalos. Suponha que a informação **Idade** seja discretizada em 4 faixas diferentes (A - até 24 anos; B - 25 à 30 anos; C - 31 à 40 anos; D - Maior que 40 anos). Assim, um discente que tenha nascido em 10 de fevereiro de 1980 terá a informação transformada para a Idade 38 anos, que após discretizada comprá a faixa C.

- Transformação: resume-se na conversão dos dados em um formato adequado, sendo compatível ao que os algoritmos de mineração de dados necessitam. Os dados são agrupados e a base de dados dá origem aos arquivos nos formatos necessários para a utilização dentro das plataformas, dentre eles:

- **CSV** (*Comma-Separated Values*) é um formato simples de armazenamento, que agrupa as informações de arquivos de texto em planilhas, para as trocas de dados com um banco de dados ou uma planilha entre aplicativos.

- **ARFF** (*Attribute-Relation File Format)* é um arquivo de texto ASCII que descreve uma lista de instâncias que compartilham um conjunto de atributos. Os arquivos ARFF possuem duas seções distintas. A primeira seção é a informação de **cabeçalho**, que é seguida da informação de **dados**. A Figura 1 apresenta um exemplo de arquivo ARFF gerado a partir da base de dados.

Figura 1: Exemplo Arquivo ARFF.

```
@relation transacoes4
@attribute COR {Parda,Branca,Negra,Amarela}
@attribute ESCOLA {Pública,Particular}
@data
Parda,Pública
Negra,Pública
Parda,Pública
Branca,Particular
Parda,Pública
Negra,Pública
Parda,Pública
Parda,Pública
Negra,Pública
Parda,Pública
Parda,Pública
Negra,Pública
Parda,Pública
Negra,Pública
```

Fonte: Autores.

A construção dos arquivos de forma correta é importante para a etapa de mineração, pois o WEKA, por exemplo, utiliza como arquivo padrão para as tarefas de mineração o formato ARFF, porém para transformar de XLS para ARFF, precisamos usar o formato CSV como intermediário.

- Mineração: é considerada a etapa mais importante do processo. Neste momento que o algoritmo escolhido é aplicado sobre os dados a fim de se descobrir padrões interessantes. Nesta fase foram utilizadas as ferramentas de mineração:
  - **RapidMiner** é uma plataforma para trabalhar com *DataScience* de forma rápida, simples e visual. Seu diferencial é a facilidade e velocidade para criar modelos preditivos, dessa forma o processo de validação e ajuste do modelo se torna mais simples.
  - **Weka** é um projeto *"Open Source*" que foi criado com o objetivo de disseminar técnicas de aprendizado de máquina. Seu diferencial é possuir um vasto arsenal de métodos e algoritmos que em conjunto com a sua interface gráfica que torna as tarefas de mineração de dados fáceis e rápidas.

- Interpretação dos Dados: nesta última etapa, os resultados obtidos até então são interpretados e avaliados, a fim de descobrir se o conhecimento adquirido será útil.

BASE DE DADOS

Os dados utilizados nos experimentos provêm do curso superior de BSI oferecido pelo IFMG - campus Sabará, e contemplam as turmas dos anos de 2013 a 2017. As informações coletadas sobre alunos estão na Tabela 1.

Tabela 1: Informações sobre os Alunos.

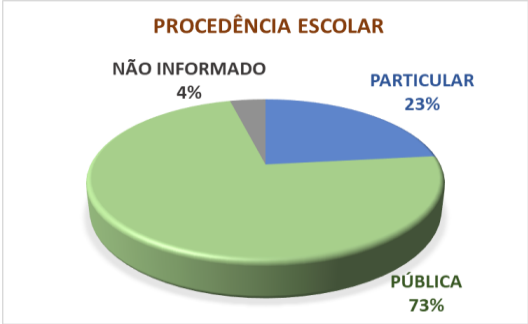
ATRIBUTOS	DESCRIÇÃO
Matrícula	Matrícula do Aluno
Procedência Escolar	Escola Pública / Escola Particular
Forma de ingresso	Vestibular /SISU
Situação	Desligado/Evasão/Regular
Status	Trancado/Cursando
Sexo	Feminino/Masculino
Cor da Pele	Branca/Negra/Parda
Data de Nascimento	Idade do Aluno
Naturalidade	Cidade Nascimento
Cidade	Cidade atual
Trabalha	Sim/Não
Deficiência	Sim/Não

RESULTADOS PARCIAIS E DISCUSSÕES

Como o projeto se encontra em desenvolvimento e devido a obtenção tardia dos dados necessários para o trabalho, ainda não foi possível descrever o impacto da mineração dos dados dos alunos. Entretanto, para realizar mineração em uma base de dados é necessário entendimento prévio do conteúdo presente na mesma. Dessa forma, alguns gráficos foram gerados com o objetivo de aprimorar o conhecimento desses dados.

A Figura 2 demonstra qual é a “Procedência Escolar” dos alunos. Conforme pode ser visto, a grande maioria dos alunos são provenientes de escolas públicas.

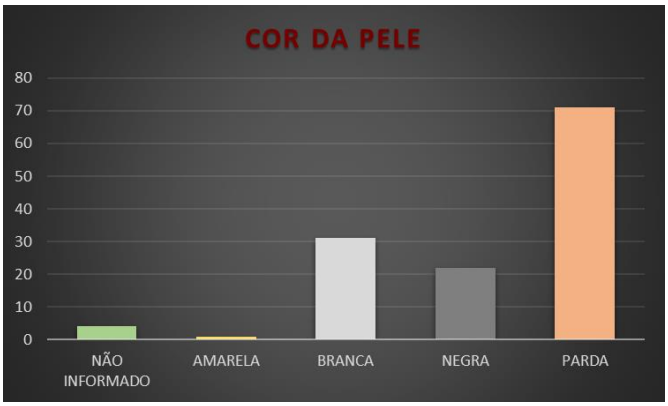
Figura 2: Diagrama de procedência escolar.



Fonte: Autores.

A Figura 3 apresenta o compilado das informações de auto declararam em relação a “Cor da Pele”. Ao analisar o gráfico, nota-se que a maior parte dos discentes se autodeclaram pardos.

Figura 3: Autodeclaração de Cor da Pele.



Fonte: Autores.

Após melhor entendimento dos dados coletados, iniciou-se o processo de mineração, buscando extrair regras de associação relevantes pertencentes à base de dados. Para isso, primeiramente utilizou-se a ferramenta *WEKA*, tendo como algoritmo selecionado para geração das regras de associação o Apriori (Agrawal, 1994). O algoritmo Apriori (Agrawal, 1994) foi proposto por *Rakesh Agrawal* e outros em 1994, sendo o primeiro algoritmo para extração de regras de associação e ainda é um dos mais utilizados atualmente.

Como teste inicial, os dados “Cor da Pele” e “Procedência Escolar” foram selecionados para execução do algoritmo. Para essa execução, o suporte mínimo foi de 12 transações. Os resultados dessa execução são mostrados na Figura 4.

A primeira regra gerada é a regra de associação: **Se cor Negra então escola Pública**. Note que a base utilizada possui 22 alunos que se auto declararam negros, sendo 95% destes provenientes de escolas públicas. Assim sendo, com 95% de confiança, pode-se inferir que o aluno negro é proveniente de escola pública.

Ainda na Figura 4, a segunda regra extraída pelo WEKA foi: **Se cor Parda então escola Pública** com 79% de confiança**.** Pela informação retornada, vê-se que do total de 71 pessoas pardas 56 são provenientes de escolas públicas.

Figura 4: Execução do Algoritmo Apriori na Ferramenta WEKA.

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 100 -T 0 -C 0.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     transacoes4
Instances:    124
Attributes:   2
              COR
              ESCOLA
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.1 (12 instances)
Minimum metric <confidence>: 0.1
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 5

Size of set of large itemsets L(2): 5

Best rules found:

  1. COR=Negra 22 ==> ESCOLA=Pública 21    <conf:(0.95)> lift:(1.26) lev:(0.03) [4] conv:(2.66)
  2. COR=Parda 71 ==> ESCOLA=Pública 56    <conf:(0.79)> lift:(1.04) lev:(0.02) [2] conv:(1.07)
```

Fonte: Autores.

É importante ressaltar que, atualmente o projeto se encontra na fase de execução de testes. Somente uma execução com os dados apresentados acima foi realizada, por isso os resultados parciais apresentados ainda são inconclusivos. Com o andamento do projeto é esperado que com testes mais detalhados encontre-se mais regras relevantes que auxiliem a coordenação e direção à nortear as ações baseadas em conhecimentos mais precisos.

CONCLUSÕES

O processo de descoberta de conhecimento em bases de dados, nos mostra verdades ocultas que fazem toda a diferença. Com resultados confiáveis, consegue-se tornar o trabalho de um administrador mais conciso, torna mais fácil de visualizar para onde e quando é realmente necessário mover recursos. O presente projeto tem como objetivo encontrar relações que sejam pertinentes ao ambiente amostral, de modo que assim sirvam como apoio à tomada de decisão aos servidores do campus. Conforme ressaltado na seção de resultados, o projeto está em fase de desenvolvimento, especificamente encontra-se na fase de execução de testes com a nova base de dados que possuem informações mais concisas e mais completa.

REFERÊNCIAS BIBLIOGRÁFICAS:

Agrawal, R.; Srikant, R.. Fast Algorithms for Mining Association Rules,Proceedings of the 20th Very Large DataBase Conference - VLDB'94, Santiago, 1994, 487-499.

Goldschmidt, R., Passos, E. “Data Mining: Um Guia Prático - Conceitos, Técnicas, Ferramentas, Orientações e Aplicações.” 1. ed. Rio de Janeiro: Editora Campus, 2005. 261 p.

Fayyad, U. M.,Piatetsky-Shapiro, G.; SMYTH, P.; Uthusumany, R.. From Data Mining to Knowledge discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996.

Silva,G.C. Mineração de Regras de Associação aplicados a dados da secretaria Municipal de Saúde de Londrina - PR. Programa de pós-graduação em Ciência da Computação, 2005.

Participação em Congressos, publicações e/ou pedidos de proteção intelectual:

O projeto foi apresentado no 3º Encontro Anual de Tecnologia da Informação do IFMG Campus Sabará (EATI), no dia 25 de maio de 2018.