APLICAÇÃO DE MINERAÇÃO DE DADOS PARA O LEVANTAMENTO DE PERFIS: ESTUDO DE CASO EM UMA INSTITUIÇÃO DE ENSINO SUPERIOR PRIVADA

Lizianne Priscila Marques **SOUTO**[1]
[1]Faculdade de Ciências Sociais e Aplicadas – FACISA, Campina Grande-PB. E-mail: lizianne.priscilla@gmail.com. Telefone: (83)8879-0679.

RESUMO

Nos últimos anos, a informação tem sido considerada como um dos recursos mais valiosos para as organizações, porém, de nada adianta ter informação se não for interpretada e utilizada, de maneira correta e segura. Nesse sentido, esse artigo sugere a utilização do KDD como uma ferramenta que possa extrair informações sobre o perfil dos alunos do curso de Sistemas de Informação, a partir de uma base de dados de uma Instituição de Ensino Superior, com o objetivo de aplicar o conhecimento obtido, no processo de tomada de decisões acadêmicas. Neste trabalho, foi realizado um estudo de caso que teve como foco, uma das tarefas de Mineração de Dados, a Clusterização, combinando informações em comum entre diversos alunos para formação de perfis.

PALAVRAS CHAVE: KDD, Mineração de Dados, Clusterização.

1 INTRODUÇÃO

Nos últimos anos pôde-se observar que as empresas utilizaram ferramentas de Tecnologia da Informação (TI) para automatizar processos e conquistar fatias cada vez maiores do mercado. O uso dessas ferramentas tem facilitado os processos de coleta e de armazenamento de dados. Com o tempo, o volume desses dados armazenados em um repositório aumenta, acumulando informações desnecessárias ou ocultando informações valiosas para a organização.

No contexto educacional, ferramentas de TI podem ajudar a fazer levantamento de informações acerca dos estudantes, de forma a apoiar decisões gerenciais. Para tal apoio é necessário conhecer a fundo o perfil dos alunos, prever antecipadamente possíveis dificuldades de aprendizagem, assim como descobrir o potencial vocacional de cada aluno, fatores essenciais para a formação acadêmica.

Os talentos devem ser devidamente explorados e o grau de dificuldade deve ser trabalhado ao longo do curso para diminuir a incerteza no direcionamento de

cada profissional ao mercado de trabalho. À medida que o número de alunos, e consequentemente, de dados a eles relacionados aumenta, diminui a capacidade de identificação dos talentos apenas de forma intuitiva, por parte dos gestores dos cursos, sendo necessário o uso de ferramentas que automatizem o processo de descoberta desse conhecimento.

Neste trabalho foi realizado o processo do KDD (Knowledge Discovery in Databases, Descoberta de Conhecimento em Bancos de Dados) como maneira de resolver o problema do excesso de dados. O KDD é um processo composto por vários passos que transformam dados brutos em informações úteis. (TAN et al., 2009).

Uma das etapas do processo de KDD é a Mineração de Dados. Existem diversas tarefas de Mineração de Dados. A escolha de uma das tarefas está relacionada com o objetivo que se deseja alcançar. Neste caso, a tarefa utilizada foi a Clusterização, onde cada *cluster* encontrado corresponde a perfis de alunos da instituição. A definição de um perfil é baseada em uma característica ou problema em particular, possibilitando a descoberta de características comuns em alunos que se enquadram no perfil.

Este estudo se faz necessário para auxiliar o coordenador a trabalhar com o aluno de maneira estratégica e diferenciada de acordo com o perfil de cada um. Esses perfis serão úteis na composição de turmas, ocupação de vagas de empregos, estágios, etc.

2 PROCESSO DE DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS (KDD)

O processo de descoberta de conhecimento em bancos de dados - *Knowledge Discovery in Databases* - KDD é uma tentativa de resolver o problema da sobrecarga dos dados estocados e sem utilidade para a organização.

Segundo Fayyad (1996 apud BOENTE et al., 2007), “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”.

O KDD é um processo amplo, composto por várias etapas, são elas: Pré-processamento, Mineração de Dados e Pós-processamento.

O pré-processamento é a primeira etapa do KDD, ela envolve todas as funções relacionadas à seleção, à organização e ao tratamento dos dados (GOLDSCHIMIDT; PASSOS, 2005). O propósito do pré-processamento é transformar dados que ainda não foram trabalhados em um formato válido para a etapa da mineração de dados (TAN et al., 2009).

“A Mineração de Dados, segunda etapa do KDD, consiste na aplicação de algoritmos específicos, que extraem padrões a partir dos dados” (FAYYAD et al., 1996 apud MACEDO; MATOS, 2010).

Ao realizar o processo de Mineração de Dados, tem-se a possibilidade de escolher dentre várias tarefas, a específica para determinada situação e com isso obter conhecimentos diversos. A tarefa está relacionada ao que se deseja encontrar nos dados (regularidades, categoria de padrões, etc.). As tarefas da Mineração de Dados são: Descoberta de Associação, Classificação, Regressão, Clusterização, Sumarização, Detecção de Desvios (*outliers*) e Descoberta de Sequências.

Dentre as tarefas de Mineração de Dados, a clusterização, foi escolhida para realizar este trabalho. Clusterização consiste em formar grupos (*clusters*) de objetos similares. De acordo com Boente et. al. (2007), essa tarefa tem como objetivo formar grupos de objetos homogênios, que compartilhem propriedades comuns e que diferenciem de elementos de outros grupos, ou seja, minimizar as diferenças entre os elementos de um *cluster* e maximizar a diferença entre os *clusters*. O K-Means foi o algoritmo utilizado para realizar a Clusterização neste trabalho. Este algoritmo é comumente utilizado e investigado, pois, implementa o método de

particionamento, bastante eficiente em classificar os dados em grupos (BOGORNY, 2003).

Na ultima fase do KDD, é realizado o Pós-Processamento. Nessa fase é possível visualizar, analisar e interpretar os resultados que foram gerados na fase de mineração de dados (GOLDSCHIMIDT; PASSOS, 2005).

Geralmente os resultados são avaliados por um especialista em KDD e o conhecimento obtido é incorporado ao sistema e apresentado ao cliente que, de acordo com seus critérios, pode-se voltar à fase inicial para novas iterações e realizar outras investigações dos dados. Na etapa de Pós-processamento são construídos gráficos, diagramas e relatórios demonstrativos para representar o conhecimento extraído no processo de *KDD* (BOENTE et al., 2007).

3 METODOLOGIA

Este estudo propõe uma metodologia para a realização de um processo de KDD, fundamentado pela realização de um estudo de caso, realizado na FACISA – Faculdade de Ciências Sociais e Aplicadas, instituição de ensino superior localizada em Campina Grande – PB, no Brasil. Os dados utilizados foram os relativos ao curso de Sistemas de Informação (SI).

Para complementar a base de dados com informações inicialmente ausentes, foi realizada uma pesquisa entre os alunos cursando a partir do segundo período do curso, a partir de um questionário contendo 14 perguntas. Essas informações diziam respeito a opiniões dos alunos acerca do curso, como seu grau de satisfação, áreas de maior interesse, o rumo que ele pretende seguir após a graduação, entre outras.

Os alunos responderam ao questionário de maneira voluntária. Atualmente existem 150 alunos (população). A amostra investigada foi de 44 alunos, ou seja, 29,33% dos alunos de SI.

Para a realização do estudo de caso, propõe-se as seguintes atividades distribuídas de acordo com o processo de KDD:

a. Pré-Processamento

Inicialmente o banco de dados (BD) deve ser explorado. Deve-se analisar todas as tabelas e seus atributos, além de, excluir do processo os dados incompletos, nulos e em branco. Em seguida, agregar aos dados selecionados e tratados, os dados obtidos na pesquisa de opnião.

b. Clusterização

Na clusterização, são executados os seguintes passos para cada Levantamento de Perfil:

1. Definição de objetivo – Inicialmente, o objetivo a ser alcançado é definido de acordo com o perfil que se deseja obter;
2. Definição do número de cluster – O número de *Clusters* (grupos) é definido pelo analista de acordo com o objetivo definido na etapa 1.
3. Escolha de atributo – Nesta etapa, o atributo chave é escolhido. A escolha do atributo chave é feita de acordo com o objetivo definido na etapa 1;
4. Mineração atributo-atributo – O atributo selecionado na etapa 3 é combinado a outro atributo e submetido ao algoritmo de mineração. Esse procedimento se repete até encontrar atributos que estejam interligados;
5. Mineração em conjunto – Os atributos que passaram pela etapa anterior são minerados, dessa vez em conjunto;
6. Resultados- Os resultados da clusterização são divulgados.

c. Pós- Processamento

Nesta fase, as informações obtidas na etapa anterior (Mineração de Dados) são analisadas e apresentadas em forma de tabela.

d. Validação

A cada obtenção de perfil, o coordenador do curso de Sistemas de Informação analisa o resultado, sugerindo mudanças, validando ou não, a

informação descoberta e acrescentando outros atributos para a investigação de novos padrões.

4 RESULTADOS E DISCUSSÃO

Neste capítulo, os resultados obtidos por meio dos Levantamentos de Perfis serão apresentados em forma de tabelas, compostas pelas colunas chamadas, atributo e *cluster*. A coluna atributo contém os dados que compõem o perfil de cada grupo (*cluster*). Nas colunas *Cluster* 1 e *Cluster* 2 contém o grupo que foi formado, nelas estão a quantidade de alunos pertencentes a cada *cluster.*

4.1 Levantamento de Perfis 1

O curso de SI é bastante procurado por pessoas do sexo masculino, o que não acontece com tanta intensidade por pessoas do sexo feminino. Como as mulheres fazem parte da minoria no curso, deseja-se saber especialmente o perfil dessas alunas, assim como a área em que elas mais se identificam.

- Objetivo: Descobrir qual área do curso que os alunos (feminino e masculino) de SI mais se identificam.

Tabela 1 – Levantamento de perfis 1.

Atributos	Cluster 1 (9 alunos)	Cluster 2 (34 alunos)
Sexo	F	M
Trabalhando	Sim, em outra área	Sim, na área de TI
Segundo_grau_escola_PubPriv	Pública	Privada
Gosta_de_programação	Não	Sim
Gosta_BancoDados	Sim	Não

Fonte: própria (2012).

Considerações: Com base no agrupamento realizado no *cluster* 1 mostrado na Tabela 1 conclui-se que, a maioria das alunas se identificam mais com a área de

Banco de Dados e não gostam de programação. Grande parte cursou o segundo grau em escola pública e trabalha em outra área diferente da área de TI.

No cluster 2 ilustrado na Tabela 1 pode-se observar que a maioria dos homens que cursam Sistemas de Informação já trabalham na área de TI e gostam de programar. Eles vêm de escola privada e, ao contrário das mulheres, não gostam da área de banco de dados.

O estudo aponta que as mulheres se identificam mais com a área de banco de dados, essa área envolve a realização de atividade como modelagens de BD e gerenciamento de dados, enquanto os homens se identificam com a área de programação, atividade que envolve raciocínio lógico e matemático.

4.2 Levantamento de Perfis 2

O curso de Sistemas de Informação oferece disciplinas de programação, banco de dados, redes de computadores, entre outras. Mas nem sempre os alunos se identificam com todas essas áreas.

- Objetivo: Deseja-se saber qual o rumo que os alunos que não gostam de programação pretendem seguir.

Tabela 2 – Levantamento de perfis 2.

Atributo	Cluster 1 (19 alunos)	Cluster 2 (24 alunos)
Gosta_de_programação	Sim	Não
Segundo_grau_escola_Pubpriv	Privada	Publica
Trabalhando	Sim, em outra área	Sim, na área de TI
Rumo_pos_graduacao	Emprego	Academia
PosGraduacao	Especialização	Mestrado

Fonte: própria (2012).

Considerações: Ao comparar os dois perfis obtidos no Levantamento de Perfis 2, observou-se que os alunos que não gostam de programar preferem tentar entrar no mestrado do que conseguir um emprego na área logo após a graduação. Isso se deve em grande parte pelo fato de que a maior parte dos empregos para recém-

formados envolve programação, o que justifica a preferência pela academia por parte dos alunos que não gostam de programar.

4.3 Levantamento de Perfis 3

O trancamento das disciplinas por parte dos alunos é um fator preocupante para a Instituição. A cada semestre vem-se observando uma grande evasão dos alunos em determinadas disciplinas. Então, surgiu a necessidade por parte da coordenação em saber quais motivos têm contribuído para isto ocorrer.

- Objetivo: Descobrir fatores que influenciam os alunos a trancar as disciplinas.

Tabela 3 – Levantamento de perfis 3.

Atributos	**Cluster 1 (24 alunos)**	**Cluster 2 (19 alunos)**
Trancou_disciplina	Sim, duas, três ou quatro	Não
Motivos_trancamento_disciplina	Dificuldade de Aprendizagem	----------
Trabalhando	Não	Sim, na área de TI
Idade	De 21 a 23 anos	De 24 a 26 anos
Rumo_pos_graduacao	Emprego	Academia
Satisfação_curso	Parcialmente Satisfeito	Completamente Satisfeito
Gosta_de_programacao	Sim	Não

Fonte: própria (2012).

Considerações: À medida que o aluno realiza a matricula em muitas disciplinas e dependendo do grau de dificuldade, ele poderá apresentar problemas de aprendizagem e a consequência futura poderá ser o trancamento de alguma(s) disciplina(s). Esse trancamento pode ocasionar o atraso da conclusão do curso, causando certa desmotivação no aluno, tornando-os parcialmente satisfeitos com o

curso. Neste estudo, observou-se que os alunos mais jovens estão mais propensos a desistir das disciplinas, estes não trabalham e teoricamente deveriam ter mais tempo para dedicar aos estudos.

É possível observar na coluna *Cluster* 2, Tabela 3, que, os alunos que não trancaram disciplinas e que estão no mercado de trabalho apresentaram maior satisfação em relação ao curso. Dessa maneira, entende-se que o curso está atingindo suas expectativas e contribuindo positivamente para executar as tarefas da profissão. É possível supor que, a partir da experiência adquirida através do mercado de trabalho e por não gostarem de programação, esses alunos gostariam de melhorar suas condições de emprego e salário, uma solução para isso seria investir mais em sua carreira tentando um mestrado.

4 CONCLUSÃO

Ter informação tornou-se uma necessidade por parte das organizações quando se trata de diferencial competitivo. Isso porque ela contribui para diminuir incertezas durante o processo de tomada de decisão. Com base nisso, as empresas têm feito uso de ferramentas de Tecnologia da Informação para facilitar a coleta e armazenamento de dados. Porém, o aumento do volume dos dados armazenados pode contribuir para ocultar informações úteis para o processo de tomada de decisão. O mesmo problema se aplica a instituições de âmbito educacional. Uma solução para este problema é o processo de KDD.

Este trabalho sugere um processo de KDD sobre os dados de uma base de alunos em uma instituição de ensino superior privada. Tal processo visa encontrar perfis de alunos baseados em problemas e dificuldades comumente encontradas no curso de Sistemas de Informação, identificando características semelhantes entre

alunos que passam por essas dificuldades e possibilitando à coordenação tomar medidas proativas para dirimi-las, num claro ganho educacional.

Foram demonstrados neste artigo três dos perfis levantados usando a metodologia de KDD proposta, cada um mostrando entre cinco e sete atributos, que representam as características em comum entre os alunos que enfrentam a mesma dificuldade. A identificação destas características fornece subsídios para a compreensão do problema como um todo, o que é necessário para que se tome qualquer medida preventiva.

A metodologia de KDD proposta não é exclusiva do curso de Sistemas de Informação, podendo ser aplicada em qualquer curso. Para tanto, é necessário identificar problemas corriqueiros do curso e fazer um levantamento de possíveis causas, de forma que o levantamento de perfis feito na base de dados do curso possa associar as causas às dificuldades, fornecendo subsídios para decisões acadêmicas.

REFERÊNCIAS

Boente, A. N. P., Goldschmidt R. R., Estrela, V. V., **Uma Metodologia de Suporte ao Processo de Descoberta de Conhecimento em Bases de Dados**, acesso em 21/07/2012, disponível em http://www.aedb.br/seget/artigos2007.php?pag=34.

Bogorny, Vania, **Algoritmos e Ferramentas de Descoberta de Conhecimento em Bancos de Dados Geográficos**, UFRGS, TI-III 1120, Research Report, 2003, pp.43, Porto Alegre.

Goldschimidt, R., Passos, E., **Data Mining: Um guia prático**. Rio de Janeiro, 2005.

Tan, P., Steinbach, M. e Kumar,V., **Introdução ao Datamining: Mineração de Dados,** Rio de Janeiro: Editora Ciência Moderna Ltda, 2009.

Macedo, D. C. e Matos, S. N., **Extração de Conhecimento Através da Mineração de Dados,** Revista de Engenharia e Tecnologia. v. 2, n. 2, p.22-30, 2010.