UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AVALIAÇÃO DE DESEMPENHO EM PROGRAMA DE FORMAÇÃO MASSIVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

MARCIA FONTES PINHEIRO

DM 27/2015

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2015

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MARCIA FONTES PINHEIRO

AVALIAÇÃO DE DESEMPENHO EM PROGRAMA DE FORMAÇÃO MASSIVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

DM 27/2015

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2015

UNIVESIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AVALIAÇÃO DE DESEMPENHO EM PROGRAMA DE FORMAÇÃO MASSIVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

MARCIA FONTES PINHEIRO

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará e julgada adequada para a obtenção do Grau de Mestre em Engenharia Elétrica na área de Computação Aplicada, elaborada sob a orientação do Prof. Dr. Ádamo Lina de Santana e coorientação do Prof. Diego Lisboa Cardoso.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2015

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Pinheiro, Marcia Fontes, 1990-
Avaliação de desempenho em programa de formação massiva utilizando técnicas de mineração de dados / Marcia Fontes Pinheiro. - 2015.

Orientador: Ádamo Lima de Santana;
Coorientador: Diego Lisboa Cardoso.
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2015.

1. Mineração de dados (computação). 2. Estratégias de aprendizagem - avaliação. 3. Ensino a distância. I. Título.

CDD 22. ed. 005.74

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AVALIAÇÃO DE DESEMPENHO EM PROGRAMA DE FORMAÇÃO MASSIVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

AUTORA: MARCIA FONTES PINHEIRO

DISSERTAÇÃO DE MESTRADO SUBMETIDA À AVALIAÇÃO DA BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA DA UNIVERSIDADE FEDERAL DO PARÁ E JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA COM ÊNFASE EM INTELIGÊNCIA COMPUTACIONAL.

APROVADA EM: 28/08/2015

Professor. Dr. Ádamo Lima de Santana – UFPA
ORIENTADOR

Professor Dr. Diego Lisboa Cardoso – UFPA
COORIENTADOR

Professor Dr. João Crisóstomo Weyl Albuquerque Costa
MEMBRO PPGEE/UFPA

Professor Dr. Marcelino Silva Da Silva
MEMBRO EXTERNO UFPA/CASTANHAL

VISTO:

Professor Dr. Evaldo Gonçalves Pelaes
COORDENADOR DO PPGEE/ITEC/UFPA

Agradecimentos

Primeiramente, agradeço a Deus por ter me dado forças e permitido que eu chegasse até aqui. À Minha mãe Roseli, agradeço de coração pelo suporte, carinho e preocupação constante. Ao meu pai Moisés, por ter acreditado e confiado em mim. À minha avó Mercês, por ser meu exemplo de esforço, luta e dedicação. Ao meu filho, Diego, pela paciência, compreensão, amor e carinho. Espero sempre poder retribuir a todo o apoio que venho tendo ao longo da minha família pela minha família. Ao meu namorado, André pelo amor, paciência, dedicação e suporte. À minha amiga Charllene, pela ajuda e encorajamento constante.

Agradeço ao meu orientador Prof. Dr. Ádamo Santana pela oportunidade de mestrado e pelos ensinamentos. Agradeço ao meu coorientador Prof. Dr. Diego pela ajuda e disponibilidade. Agradeço ao meu amigo Prof. Msc. Jacob pela ajuda, favores e encorajamentos constantes. Aos amigos do LINC, Luiz, Iago, Vincent, Igor, Gilberto, Fábio, Nathália, Paulo, Carlos pelos favores, encorajamento constante e missões a cumprir. Aos amigos do LPRAD e LEA, Eulália, Priscila, André, Dércio e Liane pela ajuda, favores e oportunidades vividas. Aos demais amigos dos laboratórios agradeço pela convivência.

Aos professores do ITA – Instituto Tecnológico da Aeronáutica e INPE – Instituo Nacional de Pesquisas Espaciais, em especial aos professores Vijay e Solon pelo tempo e conhecimento compartilhado enquanto estive em São José dos Campos, assim como aos amigos Marlon da Silva e Victor Machado.

Aos meus professores que foram verdadeiros mestres, incentivadores e motivadores profissionais para a engenheira e mestra que almejo ser.

Aos professores membros da banca examinadora Prof. Dr. Diego, Prof. Dr. João e Prof. Dr. Marcelino pelas valiosas contribuições.

Agradeço ao Governo Federal, aos Ministérios da Ciência e Tecnologia, das Comunicações e do Planejamento, Orçamento e Gestão, Secretaria de Logística e Tecnologia da Informação, Secretaria de Inclusão Digital, Assessoria de Inclusão Digital, Programa Telecentros.BR, Rede Nacional de Formação, a todos os Gestores, Coordenadores e colaboradores do Programa Telecentros.Br e aos Polos pelo amplo trabalho realizado e pela disponibilização dos indicadores e bases de dados pertencentes ao Programa Telecentros.Br, as quais essa dissertação utilizou como estudo de caso.

Finalmente agradeço à UFPA, CNPq, CAPES, PROCAD e ao PPGEE pelo suporte durante todo o mestrado e por terem me proporcionado experiências que superaram minhas expectativas.

“A verdadeira sabedoria é um dom de Deus”

Jó 28:1-28

SUMÁRIO

Lista de Figuras ..... **X**

Lista de Tabelas ..... **XII**

Lista de Abreviaturas e Siglas ..... **XIII**

Resumo ..... **XIV**

Abstract ..... **XV**

1. Introdução **1**

1.1 Definição do Problema e Motivação 1

1.2 Hipóteses e Objetivos 5

1.3 Organização do Documento 6

2 Fundamentação Teórica **7**

2.1 Considerações Iniciais 7

2.2 Mineração de Dados Educacionais 8

2.3 Mineração de Dados 10

2.3.1 Etapas de KDD 11

2.3.2 Classificação 13

2.3.3 Árvores de Decisão 13

2.3.4 Random Forest 14

2.3.5 Avaliação de Classificadores 15

2.3.6 Agrupamento 16

2.3.7 K-Means 16

2.3.8 Self-Organized Map 18

2.4 Web Mining 20

2.5 Mineração de Texto 21

2.5.1 Etapas de Mineração de Texto 24

2.6 Considerações Finais 29

3 Telecentros.Br **31**

3.1 Considerações Iniciais 31

3.2 Programa Telecentros.Br 31

3.3 Rede Telecentros.Br 32
3.3.1 Plataforma Moodle 35
3.3.2 Monitoramento d Avaliação da Formação dos Monitores 36
3.4 Considerações Finais 41

4 Trabalhos Correlatos 43
4.1 Considerações Iniciais 43
4.2 Seleção de Atributos 44
4.3 Aplicações de EDM 48
4.4 Considerações Finais 51

5 Metodologia de Avaliação de Desempenho em Programa de Formação Massiva Utilizando Técnicas de Mineração de Dados 52
5.1 Considerações Iniciais 52
5.2 Metodologia 52
5.2.1 Metodologia para Mineração de Dados Educacionais 52
5.2.2 Avaliação de Desempenho em Programa de Formação Massiva Utilizando Mineração De Dados 54
5.2.3 Perfis de Uso 60
5.2.4 Desempenho dos Alunos 61
5.2.5 Encontrar Características Educacionais 62
5.2.6 Testar Classificadores 68
5.3 Resultados 70

6 Considerações Finais 76
6.1 Contribuições 77
6.2 Publicações Geradas 77
6.3 Trabalhos Futuros 78

Referências 79

Anexo A – Artigo Aceito em Congresso 87

LISTA DE FIGURAS

Figura 2.1 – Processo de Mineração de Dados Educacionais. 9

Figura 2.2 – Etapas do processo KDD definido por Fayyad *et al.* (1996).. 12

Figura 2.3 – Representação de uma árvore de decisão. 14

Figura 2.4 –Exemplo de matriz de confusão para um problema com três classes. 16

Figura 2.5 – Clusters do algoritmo encontrados pelo K-Means. 17

Figura 2.6 – Clusters do algoritmo encontrados pelo SOM. 18

Figura 2.7 – Processo de Mineração de Texto..... 24

Figura 3.1 –Curso de Formação de Monitores do Telecentros.BR 34

Figura 3.2 – Indicadores de Participação dos Monitores. 41

Figura 4.1: Atributos para representação de estudantes..... 45

Figura 4.2: Atributos propostos para representação de estudantes..... 46

Figura 4.3 – Principais atributos para EDM segundo Romero, Ventura e García (2008) 47

Figura 4.4 –Tabela resumo contendo os atributos sobre um estudante no AVA Moodle 47

Figura 5.1 – Metodologia de Mineração de Dados Educacionais proposta. 53

Figura 5.2 – Amostra da base de observações realizadas. 63

Figura 5.3 – Distribuição das observações por avaliações realizadas. 64

Figura 5.4 – Palavras mais frequentes nas avaliações do Monitores..... 65

Figura 5.5 – Palavras mais frequentes nas avaliações do Monitores..... 65

Figura 5.6 – Amostra de observação com representação TF*IDF. 66

Figura 5.7 – Gravidez como causa de evasão da Formação. 67

Figura 5.8 – Emprego como causa de evasão da Formação. 67

Figura 5.9 – Dificuldades de acesso à Internet como causa de evasão da Formação. 68

Figura 5.10 – Amostra de dados nominais de Perfil Educacional. 69

Figura 5.11 – Amostra de dados nominais de Perfil Educacional..... 70

Figura 5.12 – Uso dos recursos pelo *cluster* K3..... 73

Figura 5.13 – Uso dos recursos pelo *cluster* K4 73

Figura 5.14 – Uso dos recursos pelo *cluster* K5 74

LISTA DE TABELAS

Tabela 4.1 –Trabalhos sobre aplicação de EDM. 48

Tabela 5.1 – Atributos utilizados padronizar 56

Tabela 5.2 – Tabelas selecionadas da base de avaliação 57

Tabela 5.3 – Resumo Perfil de Acesso..... 58

Tabela 5.4 – Perfil Educacional..... 59

Tabela 5.5 – *Clusters* encontrados por perfil de uso..... 61

Tabela 5.6 – Conceitos definidos para a classificação..... 62

Tabela 5.7 –Seis palavras mais frequentes na base de avaliações. 66

Tabela 5.8 – Avaliação de classificadores em ambiente de Formação Massiva 69

LISTA DE ABREVIATURAS E SIGLAS

AVA	Ambiente Virtual de Aprendizagem
DM	*Data Mining*
EDM	*Educational Data Mining*
IE	*Information Extraction*
IR	*Information Retrievel*
KDD	*Knowledge -Discovery in Databases*
KDT	*Knowledge Discovery from Textual Databases*
MLP	*Multi-layer perceptron*
PLN	Processamento de Linguagem Natural
SOM	*Self – Organized Maps*
TDM	*Text Data Mining*
TF*IDF	*Term Frequency – Inverse Document*
TIC	Tecnologia da Informação e Comunicação
TM	*Text Mining*
VSM	Vectorial Space Model

RESUMO

Com a evolução da aplicação de Tecnologias da Informação e Comunicação (TICs) no sistema educacional, foi fomentado o surgimento de novos métodos, técnicas e procedimentos que favoreçam a aprendizagem ativa, planejamento e gestão de cursos e suporte para superação de dificuldades no processo educacional, sejam presenciais ou a distância. Os Ambientes Virtuais de Aprendizagem (AVAs) tornaram-se fundamentais à condução de processos educacionais, propiciando a democratização da educação e permitindo a formação continuada, além de gerar grandes volumes de dados a respeito do processo de aprendizagem. Ter informações sobre o processo de aprendizagem é de extrema importância para os educadores e alunos, uma vez que permite apoiar a tomada de decisão e reflexão sobre as metodologias aplicadas no ensino, conteúdo utilizado e desempenho dos alunos. Neste sentido, esta pesquisa propõe metodologia de seleção de atributos para avaliação de desempenho de alunos de Programa de Formação Massiva utilizando técnicas de Mineração de Dados. A metodologia proposta considera identificar atributos a serem utilizados para realização de inferências relacionadas ao desempenho dos estudantes e correlacionando com aspectos sociais através de análise qualitativa e quantitativa de resultados. Esta metodologia foi desenvolvida considerando o contexto educacional e valorizando a diversidade neste processo. Para demonstrar a viabilidade da metodologia proposta aplicou-se estudo de caso em ambiente híbrido de aprendizagem massiva com bases de dados proprietárias do Programa Telecentros.BR disponibilizadas pelos gestores do Programa. No estudo de caso foi aplicada a metodologia de seleção de atributos para a mineração de dados educacionais, conseguinte foram aplicadas tarefas de classificação utilizando os algoritmos J48, *Random Forest* e *Random Tree* para predição de notas de alunos; tarefas de agrupamento utilizando os algoritmos de *K-means* para encontrar perfil de alunos baseado em *logs* de utilização do AVA e *Self-Organized Maps* (SOM) para encontrar características educacionais qualitativas a partir de avaliações qualitativas textuais. Os resultados obtidos através de estudo de caso demonstraram a viabilidade da metodologia considerando o contexto educacional e apresentam novos indicadores de desempenho aos gestores do Programa Telecentros, tais como perfil de uso do AVA, indicadores de evasão, perfil dos alunos.

Palavras-chave: Processo de aprendizagem, Ambientes Virtuais de Aprendizagem, Avaliação de desempenho, Formação Massiva, Seleção de atributos; Mineração de dados.

ABSTRACT

With the evolution of the application of Information and Communication Technologies (ICTs) in education was fostered the emergence of new methods, techniques and procedures that favor active learning, planning and management courses and support for overcoming difficulties in the educational process, be distance learning or presencial teaching. The Virtual Learning Environments (VLEs) have become fundamental to the conduct of educational processes, providing the democratization of education and enabling continuing education, as well as generating large volumes of data about the learning process. Have information about the learning process is of utmost importance for educators and students, as it allows to support decision making and reflection on the methodologies applied in education, used content and student performance. In this sense, this research proposes feature selection methodology for performance evaluation Massive Training Program students using data mining techniques. The proposed methodology considers identify attributes to be used for making inferences related to student performance and correlated with social aspects through qualitative and quantitative analysis of results. This methodology was developed considering the educational context and valuing diversity in the process. To demonstrate the feasibility of the proposed methodology was applied case study on hybrid environment of massive learning with proprietary databases from Telecentros.BR program provided by the managers of the program. In the case study was applied to feature selection methodology for Educational Data Mining, thus classification tasks were applied using the J48 algorithms, Random Forest and Random Tree to predict student grades; grouping tasks using the K-means algorithm to find profile of students based on the VLE usage logs and Self-Organized Maps (SOM) to find quality educational features from textual qualitative assessments. The results obtained through case study demonstrated the feasibility of the methodology considering the educational context and present new performance indicators to managers of Telecentros.BR program, such as profile use of AVA, evasion indicators, student profile.

Keywords: Learning Process; Virtual Learning Environments; Perfomance evaluation; Massive Training; Attribute selection; Data Mining.

1. INTRODUÇÃO

1.1 DEFINIÇÃO DO PROBLEMA E MOTIVAÇÃO

A educação passa por transformações com a inserção de TICs para realizar o processo de aprendizagem. O ensino não é mais limitado a um espaço físico, como no modelo de ensino presencial tradicional. O uso de computadores e da Internet na educação permitiram ultrapassar as paredes das escolas, propiciando aos estudantes maiores oportunidades de personalizar a educação; acessar recursos distantes; receber ajuda; ter troca de conhecimento; provocar desafios; aguçar a curiosidade; criar situações de interação mais intensas; engajar em aprender em novas formas e estimular a aprendizagem ativa.

Esta modalidade de ensino na qual o processo educacional ocorre através de tecnologias que permitem que a interação professor-aluno ocorra mesmo com separação espacial ou temporal foi denominada de Educação a Distância (EAD). Com a incorporação de TICs na educação, possibilitou –se o desenvolvimento de AVAs como novos meios de apoio a EAD, fornecendo ferramentas de troca de informações, comunicação, colaboração, interação e disponibilização de material didático. Os AVAs permitem democratizar o conhecimento, uniformizando as oportunidades educacionais e apoiando o processo de conhecimento coletivo, além de gerar rica base de dados do processo educacional.

Os AVAs armazenam todas as interações dos usuários dentro da plataforma e informações do desempenho do aluno, podendo guardar informações como: quais atividades um estudante participou, materiais que foram lidos e escritos, testes ao qual foi submetido, chats que o mesmo participou, páginas acessadas na plataforma, etc. (MOSTOW et al., 2005); assim como informações pessoais sobre usuário, como seu perfil, em sua base de dados.

Neste âmbito, ter informações sobre o processo de aprendizagem é de extrema importância para os educadores e alunos, uma vez que permite apoiar a tomada de decisão e reflexão sobre as metodologias aplicadas no ensino, conteúdo utilizado e desempenho dos alunos.

Visando investigar o processo de aprendizagem, surgiu uma linha de pesquisa que estuda as diferentes técnicas que podem ser utilizadas para extração de indicadores e conhecimento desses sistemas educacionais, denominada Mineração de Dados Educacionais (Educational Data Mining - EDM) (BARKER; ISOTANI; DE CARVALHO, 2011) (FRASCARELI; PIMENTEL, 2012) (GOTTARDO; KAESTNER; NORONHA, 2012).

EDM é uma área de pesquisa recente (FRASCARELI; PIMENTEL, 2012) e pode ser considerada uma especialização da Mineração de Dados (*Data Mining* – DM) tradicional (ROMERO; VENTURA, 2010) pois busca encontrar padrões e conhecimento não triviais em dados educacionais utilizando técnicas específicas ou técnicas clássicas de DM.

Esta pesquisa iniciou-se pela revisão da literatura e busca por problemáticas em EDM. Estudos iniciais na área aplicavam principalmente técnicas usuais de DM no domínio da educação. Ademais, verificou-se que o principal objetivo dos trabalhos envolvendo Mineração de Dados Educacionais é testar técnicas em diferentes cenários educacionais com intuito de predizer desempenho de alunos ou encontrar perfis do mesmo.

Portanto, as análises subsequentes basearam-se nesta ótica, tendo sido realizada análise das publicações mais recentes, além de teses e dissertações na área de informática e educação conforme critério da CAPES, porém, alguns trabalhos antigos se fazem notar devido à relevância para o tema e de forma a identificar possibilidades de contribuições.

Dessa forma, iniciou-se a pesquisa por revisões da literatura em EDM. O trabalho de Romero; Ventura (2007) se enquadra neste grupo. Nele os autores realizaram uma revisão dos métodos do estado da arte de EDM com destaque para Mineração na Web (*Web Mining*) e Mineração de Textos (*Text Mining - TM*) aplicados ao domínio educacional. Os mesmos autores realizaram revisão posterior (ROMERO; VENTURA, 2010) com as técnicas de DM mais utilizadas. Visando melhoria e direcionamento de futuras pesquisas, os autores apontam problemáticas que permeiam nas publicações de EDM nos aspectos de:

- Prover ferramentas de EDM mais simples de utilizar, de maneira que educadores ou usuários não especialistas em EDM possam utilizar com facilidade estas;
- Integrar ferramentas de EDM em ambientes educacionais;
- Padronizar dados e modelos, de maneira que as ferramentas EDM possam ser utilizadas em qualquer sistema educacional e disponibilizar mais bases de dados para que esta padronização possa ocorrer.

- Levar em consideração o contexto educacional na aplicação de algoritmos tradicionais de DM.

Em relação à utilização das técnicas a serem utilizadas em tarefas de EDM, Sachin; Vijay (2012) examinam o histórico e apresentam as técnicas de DM comumente utilizadas no campo educacional, como classificação, regressão, clusterização e inferência de relações. Sachin; Vijay (2012) destacaram ainda o uso de técnicas importantes como detecção de *outliers* e Mineração de Texto no campo educacional.

No que tange à avaliação de desempenho, a revisão de Hämäläinen; Vinni (2011) diferenciou-se por realizar teste quantificador de taxas de acurácia nas pesquisas de EDM utilizadas. Os autores destacam que as características particulares dos estudos resultam em variações de resultados observados, porém a média de acurácia é de 72% entre as pesquisas consideradas.

No contexto da educação no Brasil, foram encontradas duas revisões da literatura (BAKER; ISOTANI; DE CARVALHO, 2011) (RODRIGUES *et al.* (2014) sobre EDM no Brasil.

Baker; Isotani; de Carvalho (2011) apresentam que o Brasil tem potencial para promoção de EDM em benefício de milhares de alunos devido o incentivo governamental ao uso da EAD, porém os autores destacam que o cenário brasileiro necessita da disponibilização de bases de dados educacionais em larga escala para a realização da padronização e estruturação destes dados. Assim será possível o desenvolvimento de modelos que possam ser aplicados em qualquer AVA, consequentemente, será potencializando o destaque do Brasil no cenário educacional mundial através de ações que promovam o ensino eficaz nos AVAs e escolas através do uso de tecnologias educacionais que complementem o ensino.

A revisão realizada por Rodrigues *et al*. (2014) mostra a tendência crescente de utilização de EDM no cenário nacional através da modelagem e previsão de desempenho de estudantes, modelagem de grupos ou aprendizagem colaborativa, mediação ou recomendação pedagógica, apoio ao estudante e feedback, detecção ou previsão de evasão, avaliação ou modelagem do estudante, dentre outros. Os autores apontaram que o AVA Moodle é a fonte de dados mais utilizada na literatura brasileira.

As revisões acima apresentadas datam de 2007, 2010, 2011, 2012 e 2014, sendo que a revisão de Romero; Ventura (2010) tem por objetivo apresentar os desafios sobre o tema. As

outras revisões apresentam um levantamento das técnicas até então mais exploradas, testando-as em experimentos controlados. É notório que estas revisões possuem um caráter excessivamente restrito, seja no âmbito de comparação de métodos de EDM, na diversidade dos cenários educacionais etc.

Adicionalmente, estas mesmas revisões apresentam como problema à não adoção de uma padronização para experimentos envolvendo ambientes educacionais; o que aumentaria a diversidade das análises, como por exemplo a adoção de grandes bases de dados reais para fins de controle. Mas que ao mesmo tempo o processo de Mineração de Dados Educacionais obedecesse às características do contexto educacional, dessa forma correlacionado a (s) estratégia (s) aos padrões educacionais.

Outro item identificado que merece destaque diz respeito à disponibilização de bases de dados educacionais com grandes volumes de dados para testes mais amplos possibilitando a criação de modelos genéricos de aprendizagem. Em trabalhos recentes que utilizaram grandes base de dados, houve a filtragem para apenas determinado grupo de aluno, impossibilitando assim uma visão mais ampla do processo (SILVA; MORINO; SATO, 2014).

Consequentemente, viu-se nestas lacunas uma motivação para a proposição de uma metodologia de padronização de Mineração de Dados Educacionais, cobrindo a tarefa de seleção de atributos para as tarefas de Mineração de Dados e diferenciando-se por considerar o contexto educacional.

Como contribuição deste estudo tem-se a proposta de metodologia especializada de seleção de atributos com contexto educacional para a EDM já que a falta de padronização do processo de Mineração de Dados Educacionais, desde a seleção de atributos até a análise de resultados, permeia nas publicações de EDM e representa um empecilho para a plena utilização dos métodos recentemente propostos pela comunidade cientifica na academia.

A fim de comprovar a viabilidade da metodologia proposta, esta dissertação apresenta estudo de caso utilizando seleção de atributos baseada em contexto para avaliação de desempenho dos alunos do Programa Telecentros.BR a fim de correlacionar o desempenho de alunos com os perfis de uso destes no AVA Moodle. O estudo de caso busca ainda encontrar fatores socioeconômicos que possam influenciar no desempenho destes alunos.

Para a pesquisa definiu-se as seguintes hipóteses e objetivos, conforme descrito a seguir.

1.2 HIPÓTESES E OBJETIVOS

Considerando o problema exposto e as motivações deste trabalho, definiu-se a seguinte hipótese que regem esta dissertação:

- A adoção de uma padronização nos experimentos envolvendo Mineração de Dados Educacionais em AVAs aumentar a qualidade das pesquisas na área, impulsionando/propiciando a plena utilização dos métodos recentemente propostos.

Com o intuito de avaliar a hipótese, definiu-se os seguintes objetivos:

1. Identificar as tarefas que possibilitem a seleção de atributos educacionais envolvendo o contexto de forma a possibilitar a configuração de múltiplos ambientes de teste;
2. Aplicar e analisar a metodologia em estudo de caso em programa de Formação Massiva, verificando a viabilidade da metodologia proposta através da busca por perfis de uso em AVA e características contextuais influenciadoras de desempenho.

O primeiro objetivo visa abalizar o processo de desenvolvimento da metodologia, o qual deverá ser dividido em etapas sequenciais que agregarão tarefas decisórias (identificação de características do contexto educacional, interação com o AVA, etc.); para então aplicar a mineração de dados educacionais.

O segundo objetivo visa aplicar e analisar a viabilidade da metodologia proposta com dados educacionais reais de maneira que indicadores de desempenho possam ser encontrados a partir da EDM com metodologia de contexto educacional.

A avaliação de desempenho é realizada a partir de duas bases de dados disponibilizadas pelo Programa Telecentros.Br para esta pesquisa. A avaliação de desempenho possui cinco etapas:

1. Encontrar perfis de alunos a partir dos *logs* de uso do Moodle, com destaque para os recursos mais utilizados na plataforma utilizando o algoritmo para agrupamento KMeans com utilização de técnicas de DM *Web Mining*;

2. Fazer levantamento estatístico dos desempenhos dos alunos na formação através dos conceitos no Sistema de Avaliação;
3. Encontrar características educacionais a partir das observações qualitativas do Sistema de Avaliação utilizando agrupamento através do algoritmo SOM com utilização de técnicas de *Text Mining*.
4. Identificar relações entre os perfis de uso, desempenho e características educacionais dos alunos;
5. Testar classificadores baseados em Árvore de Decisão na base de avaliação para elencar potenciais métodos para predição de desempenho de alunos no processo de Formação

Com isso, espera-se possibilitar uma padronização na metodologia de pesquisas de EDM de modo a impulsionar os estudos relacionados. Outro resultado esperado é encontrar novos indicadores de desempenho no estudo de caso do Programa Telecentros.BR e retornar aos gestores do programa de Formação novas possibilidades de avaliação automática.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Este manuscrito está organizado em seis capítulos. O Capítulo 1 apresenta uma introdução à dissertação, definindo o problema a ser investigado e seus motivadores. Em seguida, tem-se o Capítulo 2, que expõe conceitos pertinentes das áreas correlacionadas com a pesquisa, mais especificamente o capítulo apresenta conceitos de Mineração de Dados. No capítulo 3, apresenta-se o Programa Telecentros.BR, que é objeto de estudo de caso dessa dissertação. Conseguinte no Capítulo 4, apresenta-se trabalhos correlatos ao tema analisado, enfatizando o estado da arte quanto a Mineração de Dados Educacionais. Capítulo 4 apresenta a metodologia proposta da dissertação e o estudo de caso. O Capítulo 5 apresenta algumas considerações acerca da pesquisa, assim como trabalhos futuros. Por fim, o Capítulo 6 apresenta as considerações finais e os próximos passos da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CONSIDERAÇÕES INICIAIS

A convergência da computação e comunicação tem produzido uma sociedade que é consumidora de informação. A IBM® estima que todo dia sejam criados 2,5 quintilhões de dados (SCHROECK *et al., 2012*). Neste âmbito, faz-se necessário o desenvolvimento de técnicas para analisar os dados a fim de extrair conhecimento útil, conferindo informações às organizações.

No campo da educação, organizações oferecem cursos com apoio de Ambientes Virtuais de Aprendizagem que armazenam uma grande quantidade de dados em registros de acesso gerados automaticamente por estes ambientes. Estes registros de uso são muito valiosos para analisar o comportamento dos estudantes e aspectos do processo de aprendizagem.

No entanto, um problema recorrente é a falta de padronização para a extração do conhecimento a partir dos dados de AVAs, haja vista que as técnicas desenvolvidas geralmente não fazem uso do contexto educacional.

Algumas técnicas de extração de conhecimento envolvem métodos estatísticos ou ainda, utilizam-se de algoritmos de aprendizado de máquina. No presente estudo utilizam-se algoritmos clássicos de aprendizado de máquina adaptados para o ambiente educacional.

Este capítulo apresentará uma breve apresentação da fundamentação teórica sobre EDM, seguido pela apresentação do processo de Mineração de Dados tradicional e suas tarefas de classificação e agrupamento. Posteriormente é feita apresentação do referencial teórico das técnicas de Mineração na Web (dados de sistemas Web) e Mineração de texto (dados textuais) que embasaram esta dissertação a utilizar para estudo de caso: i) classificação para predição de desempenho de aluno; ii) identificação de perfis de alunos baseado em agrupamento de logs de utilização do AVA;iii) e agrupamento de características qualitativas educacionais para descoberta de novos indicadores.

É importante apresentar as características e peculiaridades de cada um destes domínios de aplicação, haja vista que ambientes educacionais podem prover dados de diferentes naturezas e características, sendo necessária a adaptação ou utilização em conjunto de técnicas distintas em prol dos objetivos.

Os conceitos vistos neste capítulo servem como embasamento teórico para a metodologia e tarefas propostas nesta dissertação, visto que o estudo de caso apresenta dados provenientes de logs, linguagem natural e avaliações qualitativas em Programa de Formação Massiva, como será detalhado no Capítulo 5.

2.2 MINERAÇÃO DE DADOS EDUCACIONAIS

A Informática na Educação é uma linha de pesquisa que tem sido consolidada, como apresenta o trabalho de Frascareli; Pimentel (2012), e tem sido um tema estudado por diversos pesquisadores da área, em particular da Inteligência Artificial Aplicada à Educação (BARKER; ISOTANI; DE CARVALHO, 2011).

KDD tem sido utilizada com o intuito de investigar perguntas cientificas na área de educação como, por exemplo, quais são os fatores que afetam a aprendizagem? Como desenvolver sistemas educacionais mais eficazes? Ou ainda a relação da abordagem pedagógica e o aprendizado do aluno, estas informações podem ser úteis não somente para os educadores, mas também aos próprios alunos, uma vez que pode ser orientada para diferentes fins por diferentes participantes no processo.

Neste contexto surgiu a Mineração de Dados Educacionais, que 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. Desta maneira é possível compreender de maneira mais eficaz e adequada os alunos, como estes aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem (PINHEIRO *et al*., 2014b).

Considerando o caráter interdisciplinar, Romero; Ventura (2013) afirmam que EDM é a combinação das áreas de Computação, Educação e Estatística, como mostra a Figura 2.7. A interseção das áreas provê as três subáreas mais relacionadas com a EDM*: E-Learning*, Mineração de Dados e Aprendizado de Máquina (*Machine Learning*), e a Aprendizagem Analítica (*Learning Analytics*).

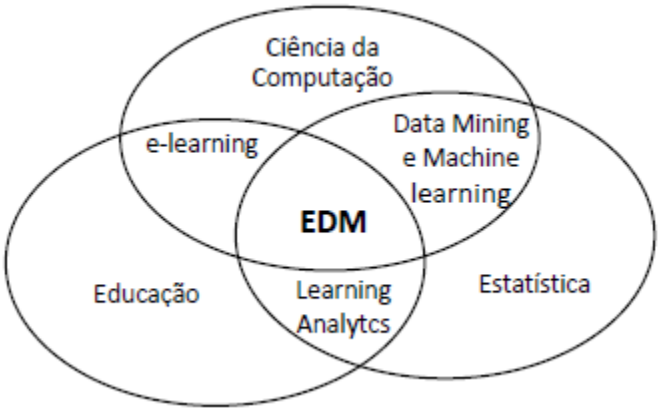

Figura 2.1 – Processo de Mineração de Dados Educacionais. Fonte: Rodrigues *et. al.*. (2014)

Neste contexto, os AVAs possibilitam a interação, troca de informações e colaboração entre os usuários. Em uma reunião realizada em Salt Lake City em 1989 pela Divisão de Estudo Independente e Telecomunicações Educacionais, foram debatidas questões como: Qual o nível de interação é essencial para uma aprendizagem eficaz? O que é uma boa interação? Como podemos alcançá-la? A interação em tempo real contribui em algo? Vale a pena o custo disso? isso é levado em consideração? (MOORE, 1989). Almejando solucionar essa problemática, Moore (1989) propôs a Teoria da Interação com o surgimento de três tipos de interação em EAD:

- Primeira Dimensão - Perfil geral de uso do AVA: dados que representam aspectos de planejamento, organização e gestão de tempo do estudante para a realização do curso. Esta dimensão utiliza indicadores gerais de quantidade e de tempo e atributos que representam atividades rotineiras e regulares dos acessos dos alunos. A interação ocorre entre o aluno e o conteúdo de estudo, a qual é essencial para a educação e resulta em mudanças na compreensão e perspectiva do aluno.
- Segunda Dimensão - Interação Estudante-Estudante: dados de interação entre os estudantes disponíveis em fóruns, chat, mensagens, etc. Esta dimensão tem como objetivo investigar a colaboração e cooperação entre os estudantes. A interação entre um aluno e outros alunos com ou sem a presença em tempo real de um professor é um desafio para os educadores, onde os membros de uma classe têm que aprender habilidades de interação em grupo

- Terceira Dimensão - Interação Estudante-Professor: dados de interação entre estudantes e professores. Esta dimensão tem como objetivo averiguar a interação entre professores ou tutores com os alunos. Na interação entre aluno e o professor estes últimos buscam estimular e manter o interesse do aluno no objeto de estudo para motivar o aluno a aprender.

Diversas técnicas são utilizadas em EDM, muitas delas originalmente são da área de Mineração de Dados, porém em grande parte são adaptadas para o domínio da educação em razão das peculiaridades dos projetos e dos dados (BARKER; ISOTANI; DE CARVALHO, 2011).

Dentre as técnicas de EDM, Barker; Isotani; De Carvalho (2011) destacam o uso de DM, *Web Mining* e *Text Mining* para: predição de notas de alunos com utilização de classificação e/ou regressão; agrupamento de perfis de alunos com base nos conteúdos e páginas visitadas; Mineração de Relações com utilização de Mineração de Regras de Associação em logs, Mineração de Correlações, Mineração de Padrões Sequenciais e/ou Mineração de Causas; Destilação de dados para facilitar decisões humanas; e Descoberta com modelos.

Nas próximas seções serão apresentados os referenciais teóricos de DM, Web Mining e *Text Mining* que embasaram o estudo de caso desta dissertação.

2.3 MINERAÇÃO DE DADOS

A quantidade de dados do mundo e das nossas vidas parece não ter fim. Computadores ubíquos permitem salvar muitos dados que antes iriam ser descartados. Discos de baixo custo e armazenamento online facilitaram o armazenamento de cada vez mais dados. Eletrônicos ubíquos gravam nossas decisões, nossas escolhas no supermercado, nossos hábitos financeiros, nossas viagens. Nossa passagem pelo mundo é gravada em bases de dados.

Podemos testemunhar uma crescente lacuna entre a geração de dados e nossa compreensão dos mesmos, dado que a compreensão das pessoas diminui de forma alarmante, à proporção que o volume de dados aumenta inexoravelmente. É estimado que a quantidade de dados armazenados duplica a cada vinte meses (WITTEN, FRANK, HALL, 2011).

Em razão desse grande volume de dados torna-se inviável a análise manual destes. Neste contexto, segundo Fayyad *et al*. (1996), a Descoberta do Conhecimento em Base de Dados (DCBD), da expressão em inglês Knowledge Discovery in Databases (KDD) surge como tentativa de solucionar o problema de sobrecarga de dados.

KDD é definido por Fayyad, et al. (1996) como um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis, tendo como vantagem a extração de conhecimento de dados sem a necessidade de conhecimento prévio ou hipóteses. Visto como um processo que deve ser automático ou semi-automático, iterativo, interativo e dividido em fases (FAYYAD *et al*., 1996) (REZENDE, 2005) (WANG *et al.*, 2005) (HAN; KAMBER, 2006) .KDD tem sido aplicada em diversas áreas do conhecimento, como por exemplo, em finanças (KOVALERCHUK; VITYAEV, 2005), bioinformática (HU, 2011), combate ao crime e terrorismo (OKONKWO; ENEM, 2011), saúde (GOSAIN; KUMAR, 2009), esportes (WICKER; BREUER, 2010), etc.

2.3.1 **ETAPAS DE KDD**

O processo de KDD é basicamente composto por um conjunto de atividades contínuas: Identificação do problema, Seleção dos dados, Pré-Processamento e Limpeza, Formatação, Escolha da Tarefa de DM, Escolha do Algoritmo de DM, Mineração de Dados, Interpretação, Consolidação e Avaliação Fayyad, et al. (1996).

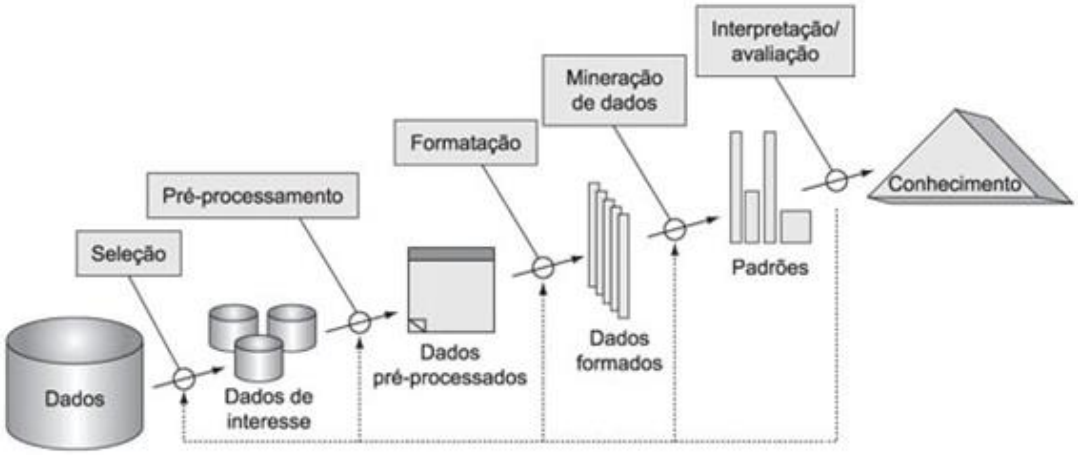

Figura 2.2 – Etapas do processo KDD definido por Fayyad *et al.* (1996**).** Fonte: Neto *et al*. (2010) apud Fayyad *et al*. (1996).

A Figura 2.1, ilustra o processo de KDD proposto por Fayyad, et al. (1996).onde a primeira etapa de KDD é a compreensão do domínio da aplicação e conhecimento a priori relevante e identificação dos objetivos do processo de KDD da visão do cliente.

A segunda etapa do processo é a criação de um conjunto de dados de interesse ou foco em um subconjunto de variáveis ou amostras de dados, nos quais a descoberta é realizada.

A terceira etapa do processo é o processo de limpeza de dados e pré-processamento. Basicamente as operações incluem a remoção de ruídos, se necessário, coletar informações necessárias para modelar ou contar ruídos, tratamento de dados ausentes, representação de séries temporais e conhecer mudanças.

A quarta atividade do processo de KDD é a redução de dados ou projeção, onde se busca encontrar características úteis para representar os dados dependendo do objetivo da tarefa. Com a redução da dimensionalidade ou métodos de transformação, o número efetivo de variáveis em consideração pode ser reduzido ou representações invariantes para os dados podem ser encontradas.

O quinto passo combina os objetivos do processo de KDD a escolha de uma determinada tarefa de Mineração de Dados, como por exemplo, classificação, associação, agrupamento, regressão ou predição.

O sexto passo é a análise exploratória e seleção de modelo e hipótese, onde se escolhe o algoritmo de Mineração de Dados a ser utilizado para a busca de padrões de dados. Esta etapa inclui a decisão de modelos e parâmetros apropriados.

O sétimo passo é a Mineração de Dados com a busca de padrões de interesse em uma forma de representação particular ou um conjunto de tais representações. Nesta etapa aplicam-se os métodos de inteligência computacional para o reconhecimento de padrões. Han; Kamber (2006) destacam os seguintes objetivos desta etapa:

2.3.2 **Classificação**

Por meio de um conjunto de atributos previsores e um atributo objetivo os quais são pertencentes a classes conhecidas, busca-se então, encontrar correlações entre esses atributos de modo a predizer a que classe cada atributo pertence. Desse modo, pode-se dizer que a tarefa de classificação tem caráter preditivo. Para avaliar a precisão de uma classificação é necessário empregar alguma técnica de medida de desempenho dentre as quais a mais empregada é a taxa de erro (ou de acerto).

No processo de classificação pode-se utilizar uma das seguintes técnicas:

2.3.3 **Árvores de Decisão**

Uma árvore de decisão parte do princípio de dividir para conquistar. Basicamente um grande problema é dividido em subproblemas aos quais recursivamente é aplicada a mesma estratégia até que uma solução seja produzida.

Dentre os algoritmos mais comuns que utilizam a idéia básica de uma árvore pode-se citar o ID3 (QUINLAN,1979), ASSISTANT (CESTNIK et Al., 1987), CART (BREIMAN et al., 1984) e C4.5 (QUINLAN,1992).

O algoritmo para a construção da árvore de decisão é guiado pelo objetivo de reduzir a entropia, o qual é a medida da aleatoriedade de uma variável aleatória e representa a dificuldade de predição da classe. A cada nó de decisão, o atributo que obteve maior ganho de informação, ou seja, o que reduziu a aleatoriedade da variável alvo é o que será escolhido para dividir os dados.

Uma árvore de decisão é um grafo acíclico direcionado no qual cada nó ou é um nó de divisão, que contém um teste condicional, ou é um nó folha.

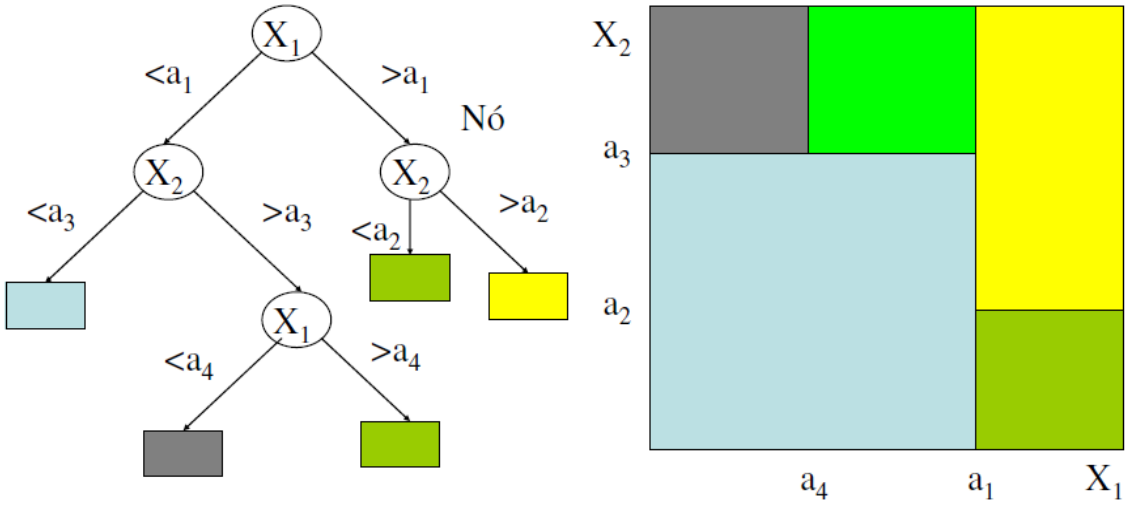

Figura 2.3 – Representação de uma árvore de decisão e as regiões de decisão no espaço de objetos.

A Figura 2.3 representa uma árvore de decisão e a divisão correspondente no espaço definido pelos atributos X_1 e X_2. A união de todas as regiões (todas as folhas) que são representadas pelas diferentes cores abrange todo o espaço de instâncias. Já a interseção das regiões abrangidas por quaisquer duas folhas é vazia.

As principais vantagens das árvores de decisão são: flexibilidade, pois faz uma cobertura exaustiva sobre o espaço de amostras e com isso pode aproximar o erro de Bayes de qualquer função; interpretabilidade, uma vez que a leitura do resultado obtido é de fácil visualização através de grafos e regras simples; e eficiência, pelo fato de sua complexidade de tempo ser linear de acordo com o número de exemplos.

Por outro lado, uma árvore de decisão não consegue lidar com valores ausentes, nem com atributos contínuos, além disso, por causa da partição de forma recursiva os dados são divididos com base no atributo de teste o que faz com que as inferências serem menos confiáveis a cada passo.

2.3.4 **Random Forest**

O algoritmo de *Random Forest* cria um conjunto de árvores de decisão no momento do treinamento do método a partir da seleção aleatória de atributos pertencentes a um vetor de

características. Após isso, calcula-se a entropia de cada atributo para que se possa separar as classes em dada posição da árvore. A saída desse classificador é a classe retornada como resposta pela maioria das árvores pertencentes à floresta.

2.3.5 **Avaliação de classificadores**

Vale ressaltar que nesta etapa são aplicadas métricas para avaliação dos algoritmos utilizados. Especificamente para a tarefa de classificação usualmente são utilizadas as métricas: taxa de erro; acurácia do classificador e matriz de confusão.

- **Erro:** Uma maneira de avaliar se um classificador f fez classificações corretas ou incorretas é através de sua taxa de erro, tal taxa mostra a proporção entre o conjunto de exemplos classificados incorretamente e é obtida pela comparação entre a classe conhecida e a classe predita pelo algoritmo ($I(y_i \neq f(x_i))$. Sua fórmula é apresentada na equação 1, onde $f(x_i)$ é a classe predita, x_i e y_i são as classes conhecidas. A taxa de erro varia entre 0 e 1 e os valores mais próximos de 0 representam os melhores resultados.

$$err(f) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq f(x_i) \quad (1)$$

- **Acurácia:** O complemento da taxa de erro corresponde à taxa de acerto e é chamada de acurácia do classificador. Sua fórmula é apresentada na equação 2. Nesse caso, os melhores valores são os mais próximos do extremo 1**.**

$$ac(f) = 1 - err(f) \quad (2)$$

- **Matriz de confusão:** Outra maneira para se observar o desempenho de um classificador é através de uma matriz de confusão. Essa matriz exibe o número de predições corretas e incorretas em cada classe, sendo que as linhas dessa matriz representam as classes verdadeiras e as colunas as classes preditas pelo classificador. Analisando o exemplo da matriz de confusão para um problema contendo 3 classes, exibidas na Figura 2.4, pode-se observar que onze das quinze amostras pertencentes à classe 1 foram corretamente classificados, sendo que um foi dito como pertencente à classe 2 e três como pertencente à classe 3. Isso nos mostra que por meio de um breve exame sobre a matriz de confusão

é possível extrair medidas quantitativas de quais classes o algoritmo tem maior dificuldade de aprendizado.

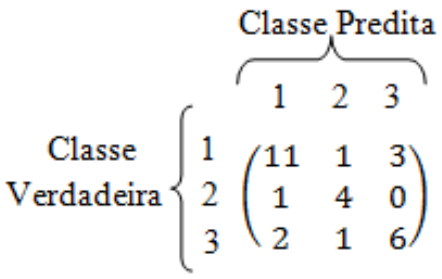

Figura 2.4 –Exemplo de matriz de confusão para um problema com três classes. Fonte: Adaptado de De Carvalho *et al*. (2011)

2.3.6 **Agrupamento**

Os algoritmos de agrupamento adotam diferentes critérios para realizar o agrupamento. Cada critério impõe uma estrutura aos dados e se os mesmos estiverem de acordo com as exigências impostas pelo critério adotado, então a estrutura de clusters pode ser encontrada. Há uma grande variedade de algoritmos de agrupamento e eles podem ser classificados de acordo com o método adotado para definir os clusters, como os hierárquicos, os particionais, os baseados em *grid* e os baseados em densidade. Dentre os algoritmos existentes podemos destacar K-Means e *Self-Oganized Maps*, os quais serão utilizados neste trabalho.

2.3.7 **K-Means**

K-Means é um método de clusterização (organização dos objetos similares, em algum aspecto, em um grupo) pertencente à classe dos algoritmos de aprendizados de máquina. Tem como finalidade dividir um determinado número de objetos em áreas chamadas de cluster. Estas áreas são constituídas por uma coleção de objetos que são similares entre si e diferentes dos objetos pertencentes aos outros grupos (*clusters).* A distância de cada objeto para cada *cluster* vai determinar em que *cluster* o objeto ficará alocado. Cada objeto pertence ao cluster no qual possui o elemento central (centróide) mais próximo deste objeto.

O algoritmo *K-Means* é iniciado com a atribuição de valores aos *k* primeiros centróides seguindo algum critério. Geralmente são escolhido os *k* primeiros pontos da tabela para cada centróide. No entanto, estes valores também podem ser gerados aleatoriamente. Depois de atribuídos os valores, há o cálculo da distância de cada objeto para cada centróide. O total de distâncias corresponde ao produto de *N* objetos por *k* centróiodes (*N* * *k*). Tendo a distância dos pontos para os centróides, os objetos são classificados de acordo com a classe a qual o objeto está mais perto do centróide, formando os clusters.

Com os grupos formados, calcula-se novamente o valor do centróide, de acordo com a posição dos objetos pertencentes ao grupo. O novo centróide é calculado através da média de cada atributo de todos os pontos pertencentes a classe. Ao final deste processo, o algoritmo volta ao segundo passo, repetindo os passos iterativamente até o ponto em que os centróides não mudem de posição, como ilustra a Figura 2.5.

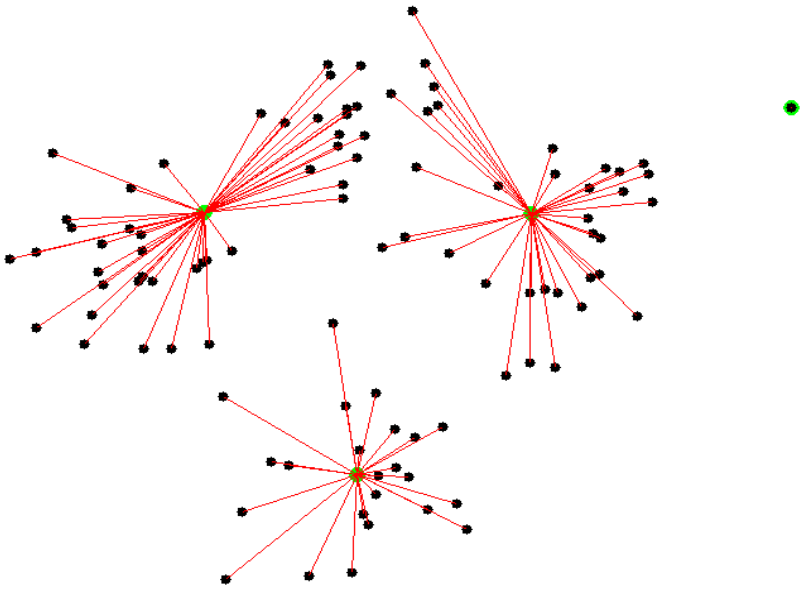

Figura 2.5 – Clusters do algoritmo encontrados pelo *K-Means*. Fonte: Shabalin (2007).

Este algoritmo pode ser usado em diversas situações. MACQUEEN (1965) afirma que as possíveis aplicações do *K-Means* estão incluídas em métodos para agrupamentos por

similaridade, previsão não linear, aproximação de distribuições multivariadas e testes não paramétricos de independência entre várias variáveis.

2.3.8 **Self-Organized Map**

Os Mapas Auto-Organizáveis (KOHONEN, 1995), também chamado de Mapas de *Kohonen*, tem a capacidade de organizar dimensionalmente dados complexos em grupos de modo a estabelecer e preservar a topologia entre os espaços de entrada e os de saída. Tem assim a capacidade de diminuir a dimensão de um grupo de dados fazendo o mapeamento do espaço original para um espaço em que está definido o arranjo dos neurônios, deste modo a representação real com relação as propriedades relevantes dos vetores de entrada são garantidas com minimização da perda de informação.

Pode ser considerada ainda como uma rede neural competitiva e com aprendizado não supervisionado, uma vez que não necessitada de um vetor de saída ou vetor alvo. No aprendizado competitivo os neurônios de saída competem entre si para serem ativados com o resultado de que apenas um neurônio de saída será ativado a cada iteração sendo este chamado de neurônio vencedor. Após isso, os neurônios são ordenados e apresentados em gráficos gradeados mapeando diferentes características de entrada formando um mapa topográfico dos padrões de entrada, como ilustra a Figura 2.5.

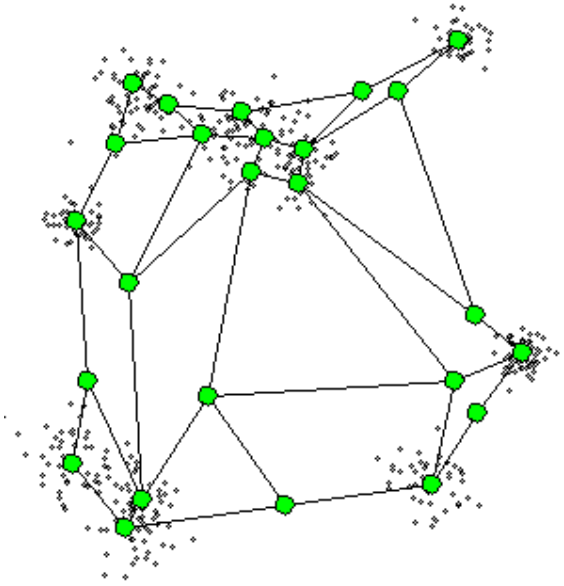

Figura 2.6 – Clusters do algoritmo encontrados pelo SOM.

Existem diversos métodos de inteligência computacional desenvolvidos para satisfazer os objetivos acima listados. Algumas destas técnicas consistem na aplicação de um determinado algoritmo de extração de padrão, outras combinam diversas técnicas visando prover uma melhor adaptabilidade e maior confiabilidade ao resultado final. Portanto, a etapa de processamento engloba a escolha do objetivo e a escolha do algoritmo (WU *et al*., 2002), culminando na extração de padrões, os quais serão avaliados e utilizados (WITTEN *et al.*, 2011). Vale ressaltar o destaque dos algoritmos Árvore de Decisão, *Random Forest*; e os algoritmos de clusterização *K-Means* e SOM.

Após a etapa de Mineração e avaliação dos algoritmos, o oitavo passo é interpretar os padrões minerados, possivelmente retornar a qualquer dos passos anteriores caso os resultados encontrados não sejam aceitáveis. Este passo pode também envolver a visualização dos padrões extraídos e modelos ou visualização dos dados.

O nono passo atua sobre o conhecimento descoberto com a consolidação e validação: usando o conhecimento diretamente, incorporando o conhecimento em outro sistema para execução de ações específicas ou documentar e relatar às partes interessadas. Esta etapa também inclui a verificação e resolução de potenciais conflitos com o conhecimento que se acreditava anteriormente.

KDD tem sido aplicada em diversas áreas do conhecimento, como por exemplo, em finanças (KOVALERCHUK; VITYAEV, 2005), bioinformática (HU, 2011), em combate ao crime e terrorismo (OKONKWO; ENEM, 2011), saúde (GOSAIN; KUMAR, 2009), esportes (WICKER; BREUER, 2010), etc. Cada domínio tem suas particularidades e características, podendo necessitar de transformações nos dados, técnicas de Mineração adaptadas ou ainda levar em consideração referencial teórico sobre o domínio no processo de KDD.

Em virtude disso, para dados educacionais, dados Web e dados textuais foi necessário realizar adaptações do processo de Mineração de Dados, resultando em processos denominados Mineração de Dados Educacionais, Mineração na Web e Mineração de Texto, respectivamente. Os processos de mineração na web e mineração de texto serão descritos de forma breve nas próximas seções, em virtude da heterogeneidade dos dados do estudo de caso realizado nesta dissertação, faz-se necessário o conhecimento dos conceitos citados.

2.4 WEB MINING

Com o explosivo crescimento do conteúdo disponível na Web tornou-se cada vez mais necessária a utilização de ferramentas automatizadas para encontrar recursos de informação desejados, acompanhar e analisar os padrões de uso. Esses fatores deram origem à Web Mining que provê informações úteis ou conhecimento de hiperlinks da Web, conteúdos de páginas e registros de uso, chamados de logs. Para explorar a mineração de informações na Web é necessário conhecer os dados a serem minerados, os quais têm sido aplicados em muitas tarefas de Mineração na Web.

A Web apresenta dados peculiares organizados de forma semi-estruturada, como as páginas *html*, e dados não estruturados, como as postagens de redes sociais e mídias por exemplo. Embora muitas técnicas de KDD possam ser aplicadas em dados provenientes da Web, diversas novas tarefas e algoritmos foram inventados na década passada visando a Mineração Web. Com base nos principais tipos de dados utilizados no processo de Mineração, as tarefas de Web Mining podem ser classificadas em três tipos:

- Mineração da estrutura da Web: tem como objetivo a descoberta de conhecimento a partir de hiperlinks os quais representam a estrutura da Web. Por exemplo, a partir dos links é possível descobrir importantes páginas Web, que, aliás, é uma tarefa chave utilizada nos motores de busca. Pode-se também descobrir comunidades de usuários que compartilham interesses comuns. A Mineração de Dados Tradicional não executa essas tarefas porque normalmente não há estrutura de links em uma tabela relacional.
- Mineração de conteúdo da Web: tem como objetivo extrair informações úteis/conhecimento de conteúdo das páginas Web. Por exemplo, pode ser realizada a classificação automática e o agrupamento de páginas Web de acordo com seus respectivos temas. Essas tarefas são semelhantes às da Mineração de Dados tradicional, porém também é possível descobrir padrões em páginas da Web para extrair dados úteis como descrições de produtos, lançamentos de fóruns, etc. Além disso, é possível extrair comentários de clientes e mensagens do fórum para descobrir sentimentos dos consumidores, tarefas não tradicionais do KDD.

- Mineração de uso da Web: tem como objetivo encontrar padrões de acesso dos usuários a partir dos logs de uso da Web, os quais são gravados a cada clique feito pelo usuário. A Mineração de uso da Web aplica muitos dos algoritmos de KDD, onde uma das questões fundamentais é o pré-processamento de dados de fluxo de cliques em *logs*, a fim de produzir dados corretos para Mineração. Esta dissertação possui forte embasamento teórico neste tipo de tarefa.

O processo de *Web Mining* é similar ao processo de KDD, diferindo na coleção de dados. No DM tradicional, os dados geralmente são coletados e armazenados em banco de dados tradicionais ou armazém de dados (*data warehouses*). Para Mineração Web, a coleta de dados é uma tarefa importante, em especial para Mineração da estrutura e conteúdo da Web, que envolve o rastreamento de um grande número de páginas Web alvo. Uma vez coletados os dados, são realizados os seguintes passos: pré-processamento, Mineração de dados Web e pós-processamento.

No campo da educação, através do uso de *logs*, *Web Mining* provê conhecimento útil para descoberta de potenciais grupos de usuários que possuam características similares permitindo que estratégias pedagógicas particulares possam ser aplicadas aos diferentes grupos com intuito de personalizar o aprendizado, incentivar potenciais alunos desistentes e estimular os alunos com metodologias mais adequadas para estilo de aprendizagem. Para esta dissertação o entendimento e aplicação de Mineração Web fazem-se necessários para levantamento das melhores técnicas a serem aplicadas em *logs* de AVAs.

Apresentados os conceitos fundamentais de Mineração na Web, a próxima seção apresentará de forma sucinta o processo de Mineração de Texto e os conceitos relevantes para esta dissertação, a qual aplicou Mineração de Texto para extração de informações educacionais a partir de avaliação qualitativa.

2.5 MINERAÇÃO DE TEXTO

A Mineração de Texto, também chamada de Mineração de Dados Textuais (*Text Data Mining* - TDM)(FELDMAN; DAGAN, 1995) ou Descoberta de Conhecimento em Bases Textuais (*Knowledge Discovery from Textual Databases* - KDT)(HEARST, 1997) consiste de

uma forma geral, no processo de extração de conhecimento a partir de coleções de documentos textuais através da identificação e exploração de padrões interessantes e não-triviais, onde estes padrões são encontrados em bases de dados textuais não estruturadas.

A MT é um campo interdisciplinar que tem como base Extração da Informação (*Information Extraction* - IE), Recuperação da Informação (*Information Retrieval* - IR), Mineração de Dados, Aprendizado de Máquina, Estatística e Processamento de Linguagem Natural (PLN).

Feldman e James (2007) definem como uma nova e empolgante área de pesquisa que tenta resolver o problema de sobrecarga de informação usando técnicas de Mineração de Dados, Aprendizado de Máquina, Processamento de Linguagem Natural, Recuperação de Informação e Gestão de Conhecimento, o qual envolve o pré-processamento de coleções de documentos, armazenamento de representações intermediárias, técnicas (análise estatística, classificação, clusterização, regras de associação, etc.) para analisar tais informações e visualização dos resultados.

É importante ressaltar que a Mineração de Texto difere de mecanismos de busca (*Search Engines*), pois neste último o usuário já sabe o que quer encontrar através da busca de informações uteis em uma coleção de documentos utilizando palavras-chave, enquanto em MT pode-se não ter conhecimento a priori (AGGARWAL; ZHAI, 2012).

O elemento chave na Mineração de Texto é o foco em coleções de texto, que pode ser qualquer agrupamento de documentos baseados em texto, sendo o documento o elemento básico na mineração de texto, o qual consiste como a unidade de dados textuais dentro de uma coleção que geralmente, mas não necessariamente, correlacionam com algum documento real como relatórios de negócios, memorandos, e-mails, pesquisas, manuscritos, artigos, comunicados de impressa, histórias, etc. (FELDMAN; JAMES, 2007).

Além disso, elementos tipográficos tais como sinais de pontuação, capitalização, números, caracteres especiais muitas vezes podem servir como uma espécie de marcadores suaves de linguagem, fornecendo pistas para ajudar a identificar importantes subcomponentes dos documentos tais como parágrafos, títulos, datas de publicação, nomes de autores, cabeçalhos e notas de rodapé. Sequência de palavras pode ser também uma dimensão estruturalmente significativa para um documento (FELDMAN; JAMES, 2007).

As operações de processamento dão suporte à Mineração de Texto na tentativa de alavancar vários elementos diferentes contidos em documentos de linguagem natural, a fim de transformá-los a partir de uma representação irregular e implicitamente estruturada em uma representação explicitamente estruturada

No entanto, dado o número potencialmente grande de palavras, frases, sentenças e elementos tipográficos que até mesmo um documento pequeno pode ter – não mencionando o vasto número de diferentes sentidos que cada um destes elementos pode ter em vários contextos e combinações – uma tarefa essencial para a maioria dos sistemas de mineração de texto é a identificação de um subconjunto simplificado de características do documento que podem ser usados para representar um determinado documento como um todo.

Tal conjunto de características pode ser visto como modelo de representação de um documento, e, os documentos individuais são representados pelo conjunto de características que seus respectivos modelos de representação contêm. Existem diversos tipos de características que podem representar um documento, porém as mais utilizadas são baseadas em caracteres, palavras, termos (palavras ou frases) e conceitos (tema).

Mesmo com as tentativas de desenvolver modelos representacionais eficientes, cada documento em uma coleção é geralmente composto de um grande número – às vezes um número extremamente grande – de características. O grande número de características necessárias para representar documentos em uma coleção afeta quase todos os aspectos da abordagem, o design e o desempenho de um sistema de TM.

Dados textuais normalmente apresentam alta dimensionalidade pois cada palavra no texto é vista como uma dimensão de maneira as tarefas de TM apresentam elevado custo computacional para processamento dos textos. Até mesmo em coleções menores, o número de palavras ou atributos necessários para representar os documentos nestas coleções pode ser extremamente grande. Por exemplo, em uma coleção de 15.000 documentos selecionados a partir de notícias, mais de 25.000 raízes morfológicas de palavras não-triviais podem ser identificadas. Mesmo quando se trabalha com tipos de características mais otimizadas, dezenas de milhares de características ainda podem ser relevantes para um único domínio de aplicação.

A alta dimensionalidade das características potencialmente representativas em coleções de documentos é um fator determinante no desenvolvimento das operações de pré-processamento de TM que visam a criação de modelos de representação mais simplificados.

Esta alta dimensionalidade também contribui indiretamente em outras condições que separam sistemas de mineração de texto de mineração de dados, tais como maiores níveis de padrões abundantes e mais características agudas para técnicas de pré-processamento.

Nas seções a seguir são apresentadas as etapas do processo de Mineração de Texto necessárias a realização de extração de conhecimento a partir de textos. Estes conceitos são importantes e embasam a tarefa de extração de características educacionais a partir de avaliações qualitativas textuais no estudo de caso realizado nesta dissertação. Tais conceitos são necessários para a aplicação de TM em dados educacionais.

2.5.1 ETAPAS DE MINERAÇÃO DE TEXTO

Nesta seção será apresentado o esquema básico do processo de Mineração de Texto com a descrição de cada etapa. O processo de MT é ilustrado na Figura 2.7 e é iniciado com a formação da base de documentos ou corpus, estes documentos passam pelo pré-processamento dos dados de forma a preparar a base de dados textual para as fases seguintes com tarefas de processamento e análise de resultados. Conseguinte, é realizado a padronização de documentos para um mesmo formato contendo informações do autor, tema, data, etc.

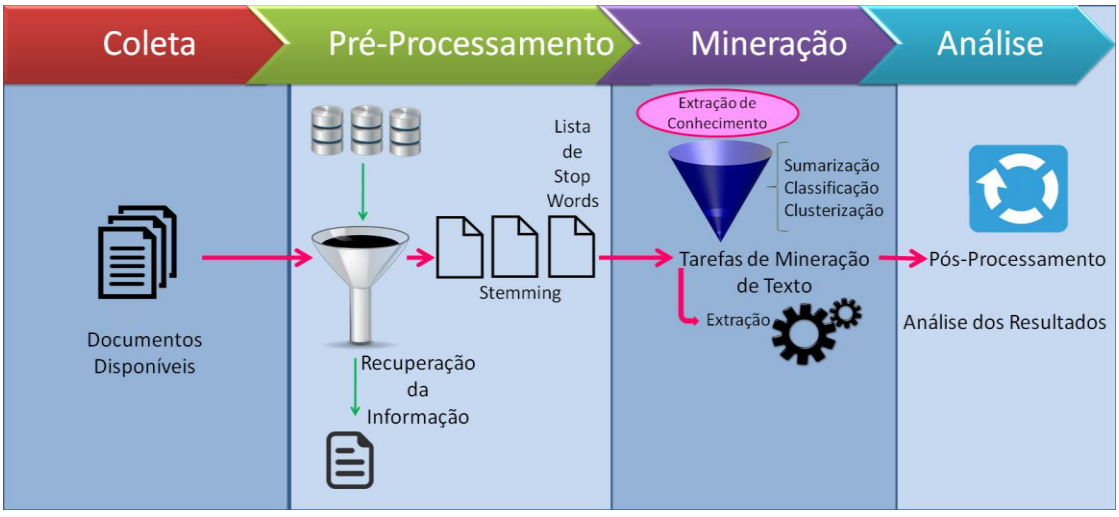

Figura 2.7 – Processo de Mineração de Texto.

Com a coleção de documentos formatada são executadas as seguintes tarefas de pré-processamento:

- Tokenization: O texto é quebrado em *tokens*, que são segmentos de texto com um significado (caracteres, palavras, frases, etc.) através de modelos de expressões regulares que eliminam caracteres especiais e retorna os *tokens* separados. A tarefa de *tokenization* é específica para um idioma, requerendo conhecimento da língua tratada na coleção de documentos.
- Remoção de *Stop Words*: Consiste na remoção de algumas palavras extremamente comuns que parecem ser de pouco valor na representação de informação em documentos. Estas palavras são chamadas de *stop words*. A estratégia geral para determinar uma lista de *stop word*s é classificar os termos de acordo com a frequência da coleção (o número total de vezes que cada termo aparece na coleção) e depois são escolhidos os termos mais frequentes, muitas vezes filtrados manualmente pelo conteúdo semântico relativo ao domínio dos documentos, como uma lista de *stop words*. Os membros desta lista de *stop words* normalmente são descartados durante a identificação de *tokens* nos documentos.
- Normalização: Algumas palavras, expressões ou valores numéricos podem ser escritos de várias formas, utilizando forma extensa, abreviaturas, etc. Para que haja ganho de informação é importante que haja a normalização de *tokens*, de modo que estes sejam correspondentes apesar das diferenças superficiais na sequência de caracteres dos *tokens*.
- *Stemming* ou *Lemmatization*: Considerando que os textos escritos em linguagem natural possuem aspectos morfológicos como temas, afixos e desinências que produzem palavras distintas, o corpus proveniente de uma coleção de documentos geralmente é muito grande devido ao número de palavras distintas. Dessa forma palavras com o mesmo significado como “trabalhar”, “trabalha” e “trabalhando” podem ser separadas como *tokens* diferentes. Objetivando diminuir o número de tipos diferentes de *tokens* em um corpus e aumentar a frequência de ocorrência de outros tipos, é necessário converter os tokens para uma forma padrão, este processo é chamado *Stemming* ou *Lemmatization*. *Stemming* reduz palavras flexionadas para sua base ou raiz (HU; LIU, 2012). Por exemplo, “trabalhar”, “trabalha”,

"trabalhará", "trabalhei" e "trabalhando" podem ser representadas como pela raiz morfológica "trabalh".

O próximo passo é realizar a transformação dos documentos em estruturas que possam ser aplicadas técnicas de KDD, neste processo as tarefas de *Information Retrieval* atuam como um filtro sobre um conjunto de documentos retornando ao usuário o resultado de um problema ou consulta particular (Lopes, 2004).

- A IR apresenta várias formas de indexar e representar os documentos, na literatura são abordados métodos de IR tradicionais e modernos, porém destaca-se o método Modelo de Espaço Vetorial (VSM) que é um dos métodos mais utilizados devido são a sua simplicidade, a facilidade que ele provê de se computar similaridades com eficiência e o fato de que o modelo se comporta bem com coleções genéricas. Faz-se necessário a apresentação deste modelo, pois este trabalho apresenta abordagem variante de VSM.Modelo de Espaço Vetorial (Vectorial Space Model – VSM): proposto por Salton e McGill (1983) utiliza representação geométrica para codificação de documentos utilizando vetores, nos quais cada componente corresponde para as diferentes palavras e o valor de cada componente reflete a frequência dos termos no documento.

  Os documentos são representados como vetores em um espaço Euclidiano N dimensional, onde cada dimensão corresponde a um termo do vocabulário. Cada termo tem um peso associado para descrever sua significância, que pode ser a frequência em um documento ou uma função dela.

  A similaridade entre dois documentos é definida como a distância entre dois pontos ou como o ângulo entre os vetores, desconsiderando o comprimento do documento. Os documentos são normalizados de forma que documentos de comprimento diferente possam ser comparados (LOPES, 2004).

  Os problemas relacionados com este tipo de representação são (LOPES, 2004): Se o vocabulário é muito grande a dimensionalidade dos vetores também será alta. Na prática, a dimensionalidade do espaço resultante é muitas vezes extremamente grande, uma vez que o número de dimensões é determinado pelo número de termos indexados distintas no corpus.

Técnicas para redução de dimensionalidade do espaço vetorial são necessárias para melhorar o processo de extração do conhecimento e reduzir o custo computacional de processamento. Os trabalhos de Jolliffee (2002), Baeza-Yates; Ribeiro-Neto (2011) e Crain *et al*. (2012) são exemplos de pesquisas na área de IR visando essa redução de dimensionalidade. As palavras são consideras por definição como independentes, sendo difícil obter uma informação semântica apenas a partir das informações textuais.

Devido à simplicidade, o modelo VSM e suas variantes são comumente usados na representação de documentos na Mineração de Texto, já que operações sobre vetores podem ser executadas de forma simples e há a existência de algoritmos de seleção de modelo, redução de dimensão e visualização de espaços em vetores (LOPES, 2004). Devido a estas e outras razões, o modelo VSM e suas variantes tem continuado em avaliações de qualidade na área da IR (BAEZA-YATES; RIBEIRO-NETO, 1999).

- Atribuição de pesos (*Weighting*): Este modelo leva em consideração a relevância dos termos no documento, cada valor numérico presente no vetor de representação VSM demonstra a relevância do termo no documento. Geralmente, valores maiores indicam maior relevância. Quando são utilizadas atribuições de pesos aos termos no vetor representativo, um cálculo de similaridade entre a consulta e o vetor representativo do documento permite recuperar itens em uma ordem de classificação de acordo com a magnitude dos coeficientes de similaridade entre a consulta e o documento (SALTON, 1983).
- Binária: A atribuição de peso binária utiliza os valores 1 (um) para identificar que o termo existe no documento e 0(zero) caso o termo não exista.
- Frequência de Termo (*Term Frequency* - TF): A abordagem baseada na Frequência de Termos leva em consideração que as características de um documento (palavras, termos, conceitos) que ocorrem frequentemente no texto dos documentos têm alguma influência sobre o conteúdo dos textos. Por isso o peso do termo t no documento d, Pesodt, deve ser igual à frequência de ocorrências, tfdt, do termo t no documento d (SALTON; MCGILL, 1983):

$$Pesodt = td*i \quad (3)$$

Normalmente ocorre a normalização para valores no intervalo [0,1] para solucionar problemas associados ao tamanho do documento, evitando-se que um termo tenha uma frequência maior simplesmente porque o documento é grande. Para cada documento divide-se o valor da frequência de cada termo pela frequência mais alta do documento considerado (LOPES, 2004).

- TF*IDF (*Term Frequency – Inverse Document Frequency*): Enquanto a abordagem baseada na Frequência de Termos não faz distinção entre os termos que ocorrem em todos os documentos da coleção e aqueles que ocorrem somente em poucos documentos. O método TF*IDF considera que a utilidade de um termo para representar um documento aumenta com a frequência do termo no documento, mas decresce com o número de documentos o qual este aparece (SALTON; MCGILL,1983). A idéia parte de que certos termos têm pouco ou nenhum poder na discriminação de relevância e deve-se usar um mecanismo para atenuar o efeito dos termos que ocorrem com muita frequência na coleção para seja significativo na determinação de relevância, para tal devemos reduzir os pesos dos termos com alta frequência na coleção total dos documentos, de maneira que o peso tfit seja reduzido por um fator que cresça com a frequência na coleção (MANNING; RAGHAVAM; SCHUTZE, 2009). Dessa forma, seja a frequência do documento dft definida pelo número de documentos na coleção que contém um termo t, o número total de documentos N na coleção, a inversa frequência de documento idf de um termo t é definida como

$$\mathrm{idf_t} = \log \frac{\mathrm{N}}{\mathrm{df_t}} \qquad (4)$$

A $\mathrm{idf_t}$ aumenta conforma a singularidade do termo entre os documentos aumenta – conforme a existência deste diminui – dando assim um peso maior ao termo (LOPES, 2014). Partindo da combinação das definições de frequência de documento e inversa frequência de documento, para produzir um peso para termo em cada documento, temos o peso tf-idf atribuído ao termo t em um dado documento d dado por

$$\text{tf-idf}_{t,d} = \mathrm{tf_{t,d}} \times \mathrm{idf_t} \qquad (5)$$

O peso $\text{tf-idf}_{t,d}$ será:

- Alto quando o termo t ocorrer várias vezes em um número pequeno de documentos, apresentado assim um alto poder discriminante nesses documentos.
- Baixo quando o termo ocorrer poucas vezes em um documento ou ocorrer em vários documentos, apresentando assim um menos acentuado sinal de relevância
- Baixo quando o termo ocorrer praticamente em todos os documentos.

Para normalizar o peso TF*IDF, é utilizada a normalização do cosseno (SALTON; WONG; YANG, 1975):

$$\frac{TF*IDF(t_k,d_j)}{\sqrt{\sum_{s=1}^{|T|}(TF*IDF(t_k,d_j))(TF*IDF(t_k,d_j))}} \quad (6)$$

Onde tk e dj são o termo e o documento em questão, respectivamente, TF*IDF(ts,dj) é a medida da coordenada de ts em dj, e |T| é o número total de termos no espaço de busca.

Esta dissertação utiliza a abordagem de representação TF*IDF, pois esta representação apresenta os termos e documentos que são altamente relevantes, além disso a codificação TF*IDF é simples, tornando-a ideal para formar a base de algoritmos mais complexos e sistemas de recuperação (RAMOS, 2003).

Por fim, as tarefas de processamento dos textos devem utilizar as técnicas mais adequadas de acordo com o problema e saída esperada, seguidas da análise de resultados obtidos do processamento.

2.6 CONSIDERAÇÕES FINAIS

Neste capítulo apresentou-se brevemente alguns temas relevantes para esta dissertação, iniciando pela contextualização de KDD, com foco no processo de extração de conhecimento

de bases de dados (FAYYAD *et al.*, 1996) – principalmente na descrição das principais etapas do processo (HAN; KAMBER, 2006).

Posteriormente, foram apresentados alguns conceitos dos processos de Mineração na Web, Mineração de Texto e Mineração de Dados Educacionais. Estes três processos são oriundos da Mineração de Dados, porém focados, respectivamente, em dados da Web, dados textuais e dados educacionais. Neste âmbito, apresentou-se o panorama das áreas, destacando conceitos relevantes como Mineração de registros de uso da Web, Teoria da Interação e representação de dados textuais.

Estas informações subsidiam o desenvolvimento de metodologias de extração de conhecimento em Ambientes Virtuais de Aprendizagem, com intuito de auxiliar o processo de aprendizagem. Romero; Ventura (2007) destacam que nestas três décadas de pesquisas, um número considerável de métodos de extração de conhecimento em dados educacionais foram propostos na literatura, com destaque para *Web Mining* e *Text Mining*. Sob esta ótica, o capítulo a seguir apresenta o Programa Telecentros.BR, o qual esta dissertação utiliza como estudo de caso para a metodologia proposta nesta pesquisa.

3 TELECENTROS.BR

3.1 CONSIDERAÇÕES INICIAIS

Inspiradas no modelo de países nórdicos (DARELLI, 2002), diversas políticas públicas tem sido implantadas nos últimos anos a fim de favorecer a inclusão digital em países da América Latina, através da criação e na utilização de centros tecnológicos comunitários, chamados também de telecentros, nos quais comunidades menos privilegiadas tem acesso público às Tecnologias de Informação e Comunicação (TIC) com computadores conectados à internet, a fim de disponibilizar espaços públicos para navegação livre e assistida, cursos e outras atividades favorecedoras de desenvolvimento local em diversas áreas (DARELLI, 2002) (SILVA *et al*., 2013) (MARTINS; FLAUZINO; DIAS, 2011) (BRASIL, 2010).

Neste contexto surge o programa Telecentros.BR, iniciativa do Governo Federal, para apoiar a implantação de novos telecentros públicos, fortalecimento de unidades já existentes no Brasil e formação massiva de agentes de inclusão digital (SILVA *et al*., 2013). O processo de Formação Massiva do Telecentros.BR foi caracterizado por ser

Neste capítulo será feita uma breve apresentação do programa Telecentros.BR que será foco do estudo de caso desta dissertação com aplicação de metodologia de Avaliação de Desempenho em Programa de Formação Massiva utilizando Mineração de Dados através de suas duas ricas bases de dados provenientes da formação utilizando a plataforma Moodle (MOODLE, 2005) e do Sistema de Avaliação de Monitores.

Para que haja a compreensão do domínio e dos dados armazenados será realizada uma breve apresentação do Telecentros.BR, das metodologias utilizadas no programa, sistemas utilizados e avaliação do processo de formação.

3.2 PROGRAMA TELECENTROS.BR

Como um item de tecnologia incentivador sócio-econômico-cultural das regiões urbanas ou rurais, os telecentros promovem ao mesmo tempo o uso de tecnologias para a cidadania e a educação tecnológica, nos quais pessoas da própria comunidade atuam como agentes de

inclusão digital e promovem o desenvolvimento social da região que atuam. (DARELLI, 2002) (SILVA *et al*., 2013) (MARTINS; FLAUZINO; DIAS, 2011).

Neste contexto, a formação de agentes de inclusão digital se torna um aspecto crucial para a efetividade dos investimentos. Neste cenário, como ação do Governo Federal, o Programa Telecentros.BR (Brasil, 2009) teve como finalidade apoiar a implantação de novos telecentros públicos, fortalecimento de unidades já existentes no Brasil e formação massiva de agentes de inclusão digital (SILVA et al., 2013).

O Programa Nacional de Apoio à Inclusão Digital nas Comunidades - Telecentros.BR foi instituído e regulamentado a partir de 2009 através de decreto e portaria interministerial no âmbito da política de inclusão digital do Governo Federal. A coordenação do programa foi realizada pelos Ministérios da Ciência e Tecnologia, das Comunicações e do Planejamento, Orçamento e Gestão. O Programa contemplou os telecentros com equipamentos de informática e mobiliário, conectividade à internet, bolsas para monitores e a formação desses monitores bolsistas e não-bolsistas (BRASIL, 2010).

As implantações e manutenções dos telecentros foram apoiadas por meio de iniciativas, programas, projetos ou ações, mas sob responsabilidade de entidades proponentes. Estas iniciativas foram selecionadas mediante seleção pública (BRASIL, 2010). Em função de melhorar a qualidade, continuar e fomentar ações promovidas pelos telecentros é constituída a Rede Telecentros.BR, em 2010, com o objetivo de realizar a formação permanente e continuada, em larga escala, dos agentes de inclusão digital - monitores (BRASIL, 2010).

3.3 REDE TELECENTROS.BR

A Rede Nacional de Formação para a Inclusão Digital – Rede Telecentros.BR – é um conjunto de atividades de qualificação de agentes de inclusão digital, nas modalidades a distância e presencial no âmbito do Programa Telecentros.BR. Como primeiro projeto da Rede de Formação, executou-se o Curso de Formação de monitores bolsistas e não bolsistas dos telecentros com agentes de inclusão digital apoiados pelo Programa Telecentros.BR, mediante a seleção de instituições habilitadas à condução do processo (BRASIL, 2010). Esses agentes devem facilitar o uso das tecnologias da informação e comunicação (TIC) como ferramentas

para alavancar transformações sociais em sua comunidade e, para isso, passam por uma formação em rede (BRASIL, 2010).

A Rede Telecentros.BR até 2011 contou com a participação de cinco pólos regionais (Norte, Nordeste, Centro-Oeste, Sudeste e Sul), dois pólos estaduais (nos estados de São Paulo e Ceará) e um Pólo Nacional. Os pólos regionais e estaduais, com apoio das iniciativas participantes do Programa Telecentros.BR, ficaram responsáveis pela formação dos agentes de inclusão digital (que são os monitores dos telecentros), tutores (responsáveis pela supervisão e acompanhamento do trabalho dos monitores), supervisores de tutoria (responsáveis pela supervisão e acompanhamento do trabalho dos tutores) e gestores (responsáveis pela administração do telecentro). O Pólo Nacional ficou responsável pela supervisão dos pólos regionais e pela coordenação pedagógica nacional do Curso de Formação dos monitores.

Para a Formação Massiva (MCAULEY *et al.*, 2010) de monitores do Telecentros.BR, os gestores definiram a utilização de recursos tecnológicos como espaços colaborativos, dentre eles a utilização do AVA Moodle para os cursos a distância de maneira que as barreiras de espaço e tempo fossem superadas na formação em larga escala. Foi proposto ainda um sistema de avaliação qualitativa dos monitores, de maneira que os tutores pudessem comentar o desenvolvimento dos monitores no processo de formação.

O Curso de Formação de Monitores do Telecentros.BR é um curso de qualificação básica que foi desenvolvido na modalidade hibrída a partir da interação do tutor com os monitores via Internet, por meio da plataforma Moodle e encontros presenciais previamente programados. As atividades do curso tiveram como referência a utilização de sistemas operacionais e softwares livres e de código aberto. O curso foi disponibilizado também de forma *off-line*, via material impresso e multimídia.

Foram promovidas também atividades presenciais, de caráter vivencial e prático, pelas iniciativas de inclusão digital dos órgãos federais, pelas iniciativas participantes do Programa Telecentros.Br e em eventos significativos de inclusão digital.

No período de fev/2010 a dez/2012, os membros dos pólos de formação da Rede Telecentros.BR se articularam para construir e aplicar o Curso de Formação de Monitores dos Telecentros e a ativação das redes sociais de agentes de inclusão social atuantes nas comunidades. Consequentemente ocasionaram as trocas de experiências e influenciaram as aplicações dos projetos comunitários nas diversas regiões do país.

O projeto de formação dos agentes de inclusão digital (monitores) foi desenvolvido em dois módulos com carga horária total de 480 horas, como mostra a Figura 1, os dois primeiros módulos disponibilizados na plataforma Moodle (REDE TELECENTROS.BR, 2013): 1) primeiro módulo com 80 horas para uma breve apresentação dos conteúdos da formação; 2) segundo módulo com foco específico no desenvolvimento de projetos comunitários e com adensamento dos eixos temáticos definidos (vide Figura 3.1), onde os monitores optam por aprofundar-se em um ou mais eixos temáticos, de acordo com a sua necessidade e/ou experiência e sem percurso pré-definido, totalizando 400 horas.

Figura 3.1 –Curso de Formação de Monitores do Telecentros.BR

Pela metodologia de ensino escolhida, a Formação da Rede Telecentros.BR tem características de formação massiva (MCAULEY et al., 2010), pois apresenta como conceitos (RIEDO et al., 2014):

- O oferecimento a público amplo que favorece justamente a amplitude geográfica, porém dependendo do acesso à rede mundial de computadores (web);
- A abertura, que pode levar à democratização do conhecimento, disponibilizando uma formação diferenciada, sem nenhum tipo de restrição de acesso tanto do ponto de vista do conhecimento prévio como econômico;

- O formato de curso, com início e fim determinados, processos avaliativos, interação entre participantes, reelaboração de conhecimentos prévios e/ou produção de novos conhecimentos.
- Tendo como modelo inovador propiciando modos alternativos para os monitores ampliarem o conhecimento, de acordo com o conteúdo escolhido a partir do próprio interesse, além de ajudar a pensar criativamente e se adaptar a paradigmas de resolução de problemas.

Sendo que o diferencial desse processo de aprendizado é a capacitação de forma cooperada para absorverem a utilização das tecnologias, a fim de contribuir na transformação social da comunidade, onde o agente de inclusão digital estava inserido.

3.3.1 **PLATAFORMA MOODLE**

O Moodle foi o AVA selecionado para o desenvolvimento do Curso de Formação de Monitores do Telecentros.BR. O curso foi estruturado em dois eixos pedagógicos: acessos a conteúdos e atividades formativas; elaboração e implementação de projetos comunitários. Ofertado em duas fases: a primeira englobava conhecer o ambiente virtual e abordagem prévia do conteúdo que foi aprofundado na fase dois que aborda principalmente o projeto comunitário, onde cada tema estudado poderá servir de apoio para realização de ações com a comunidade.

No ambiente virtual de aprendizagem Moodle, foi organizado o curso para formação dos monitores em Fase 1, Ambientação e Voo Rasante e Fase 2, Projeto Comunitário. Também foi necessário realizar um curso para formação de tutores que atuaram auxiliando e orientando os monitores durante a formação do TELECENTROS.BR. No curso dos tutores foi abordado educação a distância, articulação social e inclusão digital. Os tutores foram selecionados com base no conhecimento sobre tecnologia, valorizando a experiência em inclusão digital.

O processo de formação foi realizado em grupo de “n” monitores para 1 tutor, que fizeram o acompanhamento para o primeiro acesso a plataforma, incentivaram a participação no ambiente virtual e no desenvolvimento do projeto comunitário, assim como realizaram a avaliação dos monitores mensalmente.

Os tutores tiveram o acompanhamento e auxilio dos supervisores de tutoria no processo de formação. Os supervisores eram pessoas que fizeram parte dos pólos regionais, facilitadores

da formação, acompanhavam os tutores e monitores no processo de formação, realizaram também avaliação das turmas com base nas informações levantadas na plataforma Moodle e com as informações repassadas pelos tutores. A coordenação pedagógica realizou acompanhamento dos supervisores e tutores com base em informações obtidas na plataforma e no sistema de avaliação.

Neste contexto, vale ressaltar que a base de dados do Moodle utilizado pelo Telecentros.BR foi coletada de maneira implícita através do uso e interação dos monitores e tutores na plataforma. O Moodle possui uma base de dados relacional composta por 145 tabelas e armazena todas as interações dos usuários na plataforma em forma de logs: cada clique que o usuário realiza na plataforma para fins de navegação, bem como detalhes das atividades que os estudantes participaram.

Esta base de dados do Moodle do programa Telecentros.BR será utilizada no estudo de caso desta dissertação, haja vista a rica quantidade de dados e informações armazenadas.

3.3.2 MONITORAMENTO E AVALIAÇÃO DA FORMAÇÃO DOS MONITORES

Com intuito de avaliar o processo de formação da Rede Telecentros.BR de maneira a qualificar a estratégia implementada no Programa quanto a eficiência e eficácia através da aprendizagem dos monitores, foi estipulado um modelo de monitoramento diferenciado: (a) avaliar o desempenho do agente no curso, acompanhando suas produções, desde a Fase 1 até o desenvolvimento do projeto comunitário; (b) avaliar a formação e a contribuição para a ativação de redes de agentes de inclusão, promovendo articulações sociais em espaços para compartilhamento, resolução de problemas do dia-a-dia do telecentro e proposição articulada de projetos comunitários (SILVA *et al.*, 2013).

Visando alcançar esses objetivos, as tarefas de acompanhamento e avaliação foram divididas entre dois grupos de trabalho: grupo de avaliação pedagógica e de monitoramento gerencial.

O grupo de monitoramento gerencial propôs mecanismos que permitiram avaliar as estratégias de conectividade em rede utilizadas ao longo da formação, como mostram os trabalhos de Silva et al. (2013), de Brito et al. (2013a), de Brito et al. (2013b). O grupo de

avaliação pedagógica elaborou instrumentos de questionários, acompanhamento de projetos comunitários e avaliação da tutoria, supervisão, dentre outros. Esta dissertação explora os dados armazenados a partir do sistema de avaliação dos monitores da Rede Telecentros.BR proposto pelo grupo de avaliação pedagógica, através da utilização de técnicas de Mineração de Dados, conforme apresentado no Capítulo 4.

Como responsabilidade do Pólo Nacional da Rede de Formação, o processo de avaliação da formação dos monitores utilizou diferentes ferramentas para a construção de indicadores para revelação de informações concisas e diferentes perspectivas, levando em consideração informações e dados que pudessem expressar aspectos no âmbito do monitor.

Para monitorar características coletivas de monitores nos âmbitos das Iniciativas, dos Pólos Regionais e Nacional foram desenvolvidos indicadores que olhem coletivos de monitores agrupados de forma diferente, contribuindo para o mapeamento dos movimentos macros e que trazem à tona especificidades das culturas das instituições que os determinam.

Para melhor compreensão dos indicadores apontados pelo Telecentros.BR, esta dissertação embasou-se nos relatórios de acompanhamento mensais disponibilizados pelos gestores do Programa e o trabalho de Valentim *et al* (2011) possibilitou a ampliação da visão de quais indicadores foram utilizados no processo e quais as fontes de dados. Assim, o estudo de caso de a avaliação de desempenho para a metodologia proposta nesta dissertação pode considerar o contexto educacional através de tais indicadores.

No âmbito de coleta e registro de dados no sistema informatizado de monitoramento, o Pólo Nacional e os Pólos Regionais ficaram responsáveis por tais tarefas, bem como pela condução da avaliação formativa do cursista, como parte integrante do processo de ensino-aprendizagem. Os Polos Regionais publicavam mensalmente avaliação dos monitores para identificação de pontos que necessitavam maior atenção e manutenção da bolsa (BRASIL, 2011).

Para a obtenção dos dados supracitados foram utilizados três instrumentos: 1) Planilhas do Ministério do Planejamento; 2) Sistema de Avaliação da Formação de Monitores; e 3) Moodle.

Estes três instrumentos supracitados fornecem dados de naturezas distintas, suscitando diversidade à leitura dos resultados. Uma série histórica com os movimentos da formação foi

constituída a partir da combinação dos dados provenientes dos instrumentos e relatórios através de aspectos quantitativos.

Valentim *et al.* (2011) apresentaram que a partir AVA Moodle foram extraídos, principalmente, três dados: acessos, monitor, inserção de conteúdo em atividades − inclusos os fóruns e visualização de páginas do ambiente. Através destes dados foram gerados os seguintes indicadores com o intuito de monitorar individualmente como os monitores participaram e acompanhar o movimento dos monitores de cada iniciativa e de cada pólo (VALENTIM *et al.*, 2011):

- Monitor: Realização de Atividades; Visualização de Conteúdo;
- Iniciativas: Monitores Ativos, Taxa de Participação, Média de Atividades de Monitores; Média de Visualização de Monitores; Participação em fóruns;
- Polos Regionais: Monitores Ativos, Taxa de Participação, Média de Atividades de Monitores; Média de Visualização de Monitores; Participação em fóruns;
- Polo Nacional: Monitores Ativos, Taxa de Participação, Média de Atividades de Monitores; Média de Visualização de Monitores; Participação em fóruns;

As Planilhas do Ministério do Planejamento informam a entrada e exclusão de monitores do Programa em períodos mensais tendo como fonte desses dados a equipe do Programa Nacional de Apoio à Inclusão Digital nas Comunidades – Telecentros.BR que atua no Ministério. Estas planilhas fornecem os seguintes dados dos Monitores: nome, CPF, email, iniciativa a que pertencem cidade de atuação, UF da Cidade e situação do Monitor (convocado ou excluído) (VALENTIM *et al.*, 2011).

O Sistema de Avaliação de Monitores foi utilizado para obtenção de dados do acompanhamento que os tutores fizeram dos monitores, com fonte de dados oriunda dos tutores atuantes nos Pólos Regionais no curso da Formação. Os dados obtidos do sistema são (VALENTIM *et al.*, 2011):

- Avaliação da realização de atividades por Monitor – Tutor;
- Interesse do monitor na formação – Visão do Tutor;
- Compreensão de formação pelo monitor – Visão do Tutor;
- Número de atualização mensais em redes sociais;

- Observações Gerais.

Este trabalho explora os dados do Sistema de Avaliação da Formação de Monitores, cujo foco está na avaliação qualitativa do ensino-aprendizagem dos monitores, onde cada avaliação é composta de dezesseis questões sendo onze questões objetivas fechadas e cinco questões subjetivas abertas considerando o envolvimento do monitor e produção na formação, sua compreensão e interesse nos conteúdos e atividades, contatos que realiza com o seu tutor, participação em eixos temáticos e projetos, uso de redes sociais e aspectos colaborativos. As respostas às perguntas subjetivas abertas, chamadas de observações, complementam a avaliação sob os pontos analisados, pois essas respostas apresentam o parecer do tutor em relação a cada tópico abordado na avaliação do monitor, dessa maneira, apresentam informações valiosas em relação ao aprendizado, evasão e participação da Formação.

De posse dos indicadores citados, o Pólo Nacional da Rede de Formação produzia os seguintes produtos de monitoramento e avaliação da Formação: Matriz de Acompanhamento dos Monitores – Iniciativas; Matriz de Acompanhamento dos Monitores – Pólos Regionais; Relatórios de Monitoramento do Curso por Iniciativa; e Relatórios de Monitoramento do Curso por Pólo Regional.

A Matriz de Acompanhamento é oriunda da integração dos dados provenientes das Planilhas e dos questionários do Sistema de Avaliação, resultando em uma matriz representante do monitoramento individual dos monitores de cada Iniciativa. A matriz tem como principal objetivo informar a presença dos monitores na Formação e é composta pelos seguintes dados (VALENTIM *et al*., 2011):

- Monitor – CPF
- Monitor – Nome
- Tutor – CPF
- Visualização de Conteúdo
- Realização de Atividades – Moodle
- Contatos com o Monitor

A Matriz de Acompanhamento dos Monitores – Pólos é oriunda da integração dos dados provenientes das Planilhas, do Moodle e dos questionários do Sistema de Avaliação, resultando

em uma matriz representante do monitoramento individual dos monitores do Polo. A matriz tem como principal objetivo informar a presença dos monitores na Formação e é composta pelos seguintes dados (VALENTIM *et al*., 2011):

- Monitor – CPF
- Monitor – Nome
- Tutor – CPF
- Visualização de Conteúdo – Moodle
- Realização de Atividades – Moodle
- Realização de Atividades – Tutor
- Relação Monitor – Formação
- Intensidade de Publicações em Redes Sociais
- Contato com o Monitor

E por fim, os relatórios de monitoramento do curso por Iniciativa expressam o comportamento dos monitores de cada iniciativa em relação ao curso, permitindo que as Iniciativas possam realizar um acompanhamento mensal de como está sendo a Formação para os monitores e reavaliar as estratégias de gestão em relação a estes e seus telecentros, de acordo com o número de monitores convocados, monitores ativos e taxa de participação, por exemplo, como ilustra o trecho de relatório na Figuras 3.2 Estes relatórios têm como fonte os dados do Moodle e das Planilhas de acompanhamento (VALENTIM *et al*., 2011):

- No âmbito da Iniciativa: Monitores Convocados, Monitores Ativos, Taxa de Participação, Média de Atividades de Monitores, Média de Visualização de Monitores e Participação em Fóruns.
- No âmbito Nacional: Monitores Convocados, Taxa de Participação, Monitores Ativos, Média de Atividades de Monitores, Participação em Fóruns.

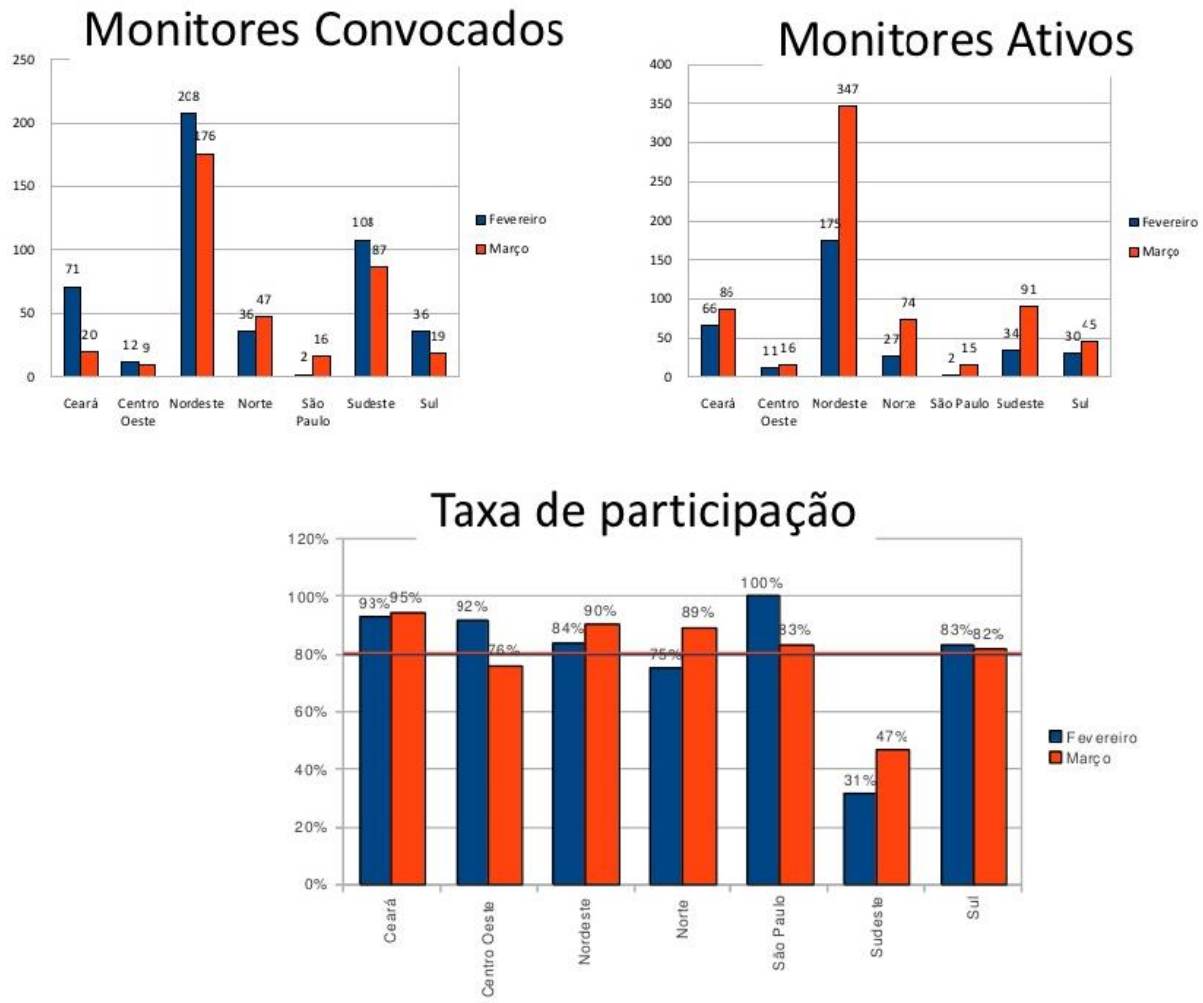

Figura 3.2 – Indicadores de Participação dos Monitores. Fonte: Rede de Formação Telecentros.Br (2011a)

3.4 CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado o projeto Telecentros.BR contemplando uma visão ampla da abordagem de inclusão digital e social desenvolvida por universidades federais, movimentos sociais e órgãos do Governo Federal (Ministérios). E ainda pela colaboração das entidades responsáveis por implantar os telecentros nas comunidades.

Destaca-se o uso da plataforma Moodle no processo de formação de monitores e tutores, a participação dos supervisores e coordenação pedagógica que na formação da Rede

trabalharam em colaboração para desenvolverem os recursos didáticos, envolvendo o conteúdo e o design instrucional do curso.

A Rede de Formação estruturou o curso, realizou acompanhamento do processo de inclusão digital e social junto aos monitores, através de dados gerados pela plataforma Moodle e pelo Sistema de Avaliação qualitativa de Monitores, realizada pelos tutores, supervisores e coordenações pedagógicas de cada pólo.

As bases do Moodle e do Sistema de Avaliação serão utilizadas como estudo de caso de Avaliação de Desempenho de Formação Massiva utilizando Mineração de Dados.

No próximo capítulo, serão abordados os trabalhos correlatos referentes à abordagem realizada nesta pesquisa.

4 TRABALHOS CORRELATOS

4.1 CONSIDERAÇÕES INICIAIS

No campo da educação, a utilização de AVAs proporciona novas possibilidades de aprendizagem na metodologia de ensino, armazenando todas as interações dos usuários dentro da plataforma através de logs. Como área de pesquisa em crescimento (Frascareli e Pimentel, 2012), diversas pesquisas têm sido feitas no campo da EDM utilizando logs de AVAs e tem sido um tema estudado por diversos pesquisadores da área, em particular da Inteligência Artificial aplicada à Educação (BARKER; ISOTANI; DE CARVALHO, 2011).

Como exposto no Capítulo 2, existem diversas classes de métodos de EDM, muitos deles originados da Mineração de Dados, Mineração da Web e Mineração de Texto, porém em grande parte são adaptadas para o domínio da educação em razão das peculiaridades dos projetos (BARKER; ISOTANI; DE CARVALHO, 2011).

Dessa forma, o presente capítulo apresenta um panorama dos estudos relacionados ao tema, explicitando os trabalhos julgados mais relevantes, primeiramente, de forma genérica – com um hiato para destacar a seleção de atributos, métodos utilizados; para posteriormente, restringir-se aos estudos relacionados ao Programa Telecentros.Br, estudo de caso deste trabalho.

A pesquisa por trabalhos relacionados à EDM deu-se, inicialmente, pela busca na literatura por trabalhos que apresentassem uma síntese dos métodos até então desenvolvidos e que aplicassem EDM em logs de AVAs. A pesquisa bibliográfica foi focada em artigos recentes de eventos científicos, principais periódicos, além de teses e dissertações na área de informática e educação conforme critérios da CAPES, porém alguns trabalhos mais antigos se fazem notar devido a relevância para o tema.

A partir das revisões de literatura encontradas (ROMERO; VENTURA, 2007) (ROMERO; VENTURA, 2010) (HÄMÄLÄINEN; VINNI; 2011) (BAKER; ISOTANI; DE CARVALHO, 2011) (SACHIN; VIJAY, 2012) (RODRIGUES *et al.*, 2014) buscou-se por trabalhos que propusessem metodologia de padronização de EDM, seleção de atributos ou aplicações a fim de encontrar o estado da arte e lacunas de contribuição.

4.2 SELEÇÃO DE ATRIBUTOS

Deixando os trabalhos de revisão de literatura para imersão em trabalhos que propõem seleção de atributos para tarefas de EDM, percebe-se que as dimensões de interação propostas por Moore (1989) são utilizadas por alguns trabalhos na etapa de seleção de atributos de forma empírica.

Gottardo, Kaestner e Noronha (2012a) propuseram um conjunto de atributos genéricos da base de dados do AVA Moodle, mostrados na Figura 4.1, a serem utilizados em tarefas de EDM. O atributo "Resultado_Final" representa a classe objetivo que era originalmente representada por valores contínuos e foi transformada para valores categóricos. Os autores utilizaram procedimento de transformação neste atributo utilizando o algoritmo de discretização não-supervisionada *equal-width*, disponível na ferramenta *Weka*, que divide o intervalo de valores possíveis em subintervalos de mesmo tamanho.

Conseguinte, Gottardo; Kaestner; Noronha (2012a) aplicaram a tarefa de classificação a fim de predizer o desempenho dos estudantes a partir dos atributos selecionados. Os autores tiveram como objetivo testar a eficácia dos métodos de classificação Random Forest e Multilayer Perceptron para diferentes cenários para os atributos selecionados: atributos com valores em sua forma original e valores numéricos, assim como experimentos para apenas alguns atributos selecionados.

Os autores apontaram a viabilidade da utilização de um conjunto amplo de atributos para representação de estudantes, potencialmente generalizáveis a diversos cenários de cursos EAD, dados os resultados dos experimentos realizados, porém os autores não propuseram metodologia de seleção de atributos e a utilização de dimensões de interação para esta tarefa, bem como a consideração do contexto educacional dos alunos.

Dimensão	Atributo	Descrição
Perfil Geral de uso do AVA	nr_acessos	Número total de acesso ao AVA
	nr_posts_foruns	Número total de postagens realizadas em fóruns
	nr_post_resp_foruns	Número total de respostas postadas em fóruns referindo-se a postagens de outros participantes (estudantes, professores, tutores)
	nr_post_rev_foruns	Número total de revisões em postagens anteriores realizadas em fóruns
	nr_sessao_chat	Número de sessões de chat que o estudante participou
	nr_msg_env_chat	Número de mensagens enviadas ao chat
	nr_questoes_resp	Número de questões respondidas
	nr_questoes_acert	Número de questões respondidas corretamente
	freq_media_acesso	Frequência média em que o estudante acessa o AVA
	tempo_medio_acesso	Tempo médio de acesso ao sistema
	nr_dias_prim_acesso	Número de dias transcorridos entre o início do curso e o primeiro acesso do estudante no AVA
	tempo_total_acesso	Tempo total conectado no sistema
Interação Estudante-Estudante	nr_post_rec_foruns	Número de postagens do estudante que tiveram respostas feitas por outros estudantes.
	nr_post_resp_foruns	Número de respostas que o estudante realizou em postagens feitas por outros estudantes.
	nr_msg_rec	Número de mensagens recebidas de outros participantes durante a realização do curso.
	nr_msg_env	Número de mensagens enviadas a outros participantes durante a realização do curso.
Interação Estudante-Professor	nr_post_resp_prof_foruns	Número de postagens de estudantes que tiveram respostas feitas por professores ou tutores do curso
	nr_post_env_prof_foruns	Número de postagens de professores ou tutores que tiveram respostas feitas por estudantes
	nr_msg_env_prof	Número de mensagens enviadas ao professor/tutor durante a realização do curso.
	nr_msg_rec_prof	Número de mensagens recebidas do professor/tutor durante a realização do curso.
Objetivo da Previsão	resultado_final	Resultado final obtido pelo estudante no curso. Representa classe objetivo da técnica de classificação.

Figura 4.1: Atributos propostos por Gottardo; Kaestner; Noronha (2012a) para representação de estudantes

Outro trabalho que realizou seleção de atributos com base nas indicações de Moore (1989) foi o de Santana, Maciel e Rodrigues (2014) que optaram por utilizar a dimensão "perfil de uso do AVA" e selecionaram os atributos, mostrados na Figura 4.2, para representação de estudantes em um AVA.

Dimensão	Atributo	Representação
Perfil de Uso do AVA	Desempenho Final	Result_final
	Número total de acesso ao fórum	Sum_int_forum
	Número total de interações com as vídeo-aulas	Sum_int_video
	Número total de interações com o material da disciplina (Caderno)	Sum_int_mat
	Número total de interações com as apresentações em Slides.	Sum_int_ppt
	Tempo médio de acesso no ambiente	Media_acesso

Figura 4.2: Atributos propostos por Santana, Maciel e Rodrigues (2014) para representação de estudantes

Para teste dos atributos, os autores utilizaram os atributos selecionados em base de dados de um curso EAD contendo uma população de 79 estudantes, desconsiderando os alunos desistentes. Santana, Maciel, Rodrigues (2014) realizaram dois experimentos: 1) discretização e inserção das classes para alunos com maiores notas (A), menores notas(C), demais alunos (B); 2) discretização e inserção das classes Aprovado ou Reprovado, para alunos com notas maiores ou iguais a 70 e menos que 70, respectivamente.

O objetivo dos autores foi testar o desempenho dos classificadores *Random Forest*, MLP, NaiveBayes, SVM, KNN, J48 e RBF da ferramenta Weka para os atributos selecionados. Apesar de utilizar dimensões de interação para seleção de atributos, este trabalho executou tal tarefa de forma empírica e sem metodologia de padronização de aplicação de EDM.

Romero; Ventura; Garcia (2008) destacam os principais atributos que podem ser utilizados para EDM, como mostra a Figura 4.3, porém os autores indicam que é necessário criar uma nova tabela capaz de resumir a informação que é necessária para o objeto de estudo, como mostra a Figura 4.4 uma nova tabela que concentra toda a informação sobre um estudante.

Name	Description
mdl_user	Information about all the users.
mdl_user_students	Information about all students.
mdl_log	Logs every user's action.
mdl_assignement	Information about each assignment.
mdl_assignment_submissions	Information about assignments submitted.
mdl_chat	Information about all chatrooms.
mdl_chat_users	Keeps track of which users are in which chatrooms.
mdl_choice	Information about all the choices.
mdl_glossary	Information about all glossaries.
mdl_survey	Information about all surveys.
mdl_wiki	Information about all wikies.
mdl_forum	Information about all forums.
mdl_forum_posts	Stores all posts to the forums.
mdl_forum_discussions	Stores all forums' discussions.
mdl_message	Stores all the current messages.
mdl_ message_reads	Stores all the read messages.
mdl_quiz	Information about all quizzes.
mdl_quiz_attempts	Stores various attempts at a quiz.
mdl_quiz_grades	Stores the final quiz grade.

Figura 4.3 – Principais atributos para EDM segundo Romero, Ventura e García (2008)

Name	Description
course	Identification number of the course
n_assigment	Number of assignments handed in.
n_quiz	Number of quizzes taken.
n_quiz_a	Number of quizzes passed.
n_quiz_s	Number of quizzes failed.
n_messages	Number of messages sent to the chat.
n_messages_ap	Number of messages sent to the teacher.
n_posts	Number of messages sent to the forum.
n_read	Number or forum messages read
total_time_assignment	Total time spent on assignment.
total_time_quiz	Total time used in quizzes.
total_time_forum	Total time used in forum.
mark	Final mark the student obtained in the course.

Figura 4.4 –Tabela resumo contendo os atributos sobre um estudante no AVA Moodle segundo Romero, Ventura e García (2008)

Romero; Ventura; Garcia (2008) também embasaram estudo nas indicações de Moore (1989), especificamente na dimensão de "perfil de uso do AVA", porém os autores realizaram a seleção de atributos sem metodologia padronizada considerando apenas aspectos sociais.

Os trabalhos correlatos supracitados enfatizaram a falta de metodologia de seleção de atributos em EDM e aplicação de acordo com cenários empíricos testados. Dessa forma, justifica-se a padronização de seleção de atributos em EDM pois utilizando a mesma metodologia, trabalhos poderão ser comparadas e assim haverá uma evolução nas pesquisas em busca de novos métodos e técnicas.

4.3 APLICAÇÕES DE EDM

Partindo dos trabalhos que realizam seleção de atributos e adentrando o campo de aplicações de EDM percebe-se que o estudo dos diversos métodos de EDM é uma tarefa antiga (SANJEEV; ZYTKOW, 1995) (ZAÏANE ET AL, 1998.) (BECK; WOOLF, 2000).

Atualmente, boa parte dos estudos que desenvolvem esta linha de pesquisa têm em vista contribuições ao domínio da aplicação. Desta forma, analisou-se 14 trabalhos publicados, apresentados na Tabela 4.1 de forma a analisar as aplicações de EDM de acordo com o objetivo educacional e a tarefa utilizada no processo.

Tabela 4.1 –Trabalhos sobre aplicação de EDM.

Autores	Objetivo Educacional	Tarefa	Fonte de Dados
DRINGUS (2005)	Modelagem de Grupos ou Aprendizagem Colaborativa	Mineração de Texto	Allaire's Cold Fusion
ROMERO; VENTURA; GARCIA (2008)	Modelagem de Grupos / Estimativa de desempenho de estudante	Mineração de dados	Moodle
BEER; CLARK; JONE (2010)	Indicadores de engajamento	Estatística	Blackboard / Moodle
MACFADYEN; DAWSON, (2010)	Avaliação ou Modelagem do Estudante	Regressão	Moodle
AZEVEDO; BEHAR; REATEGUI (2011)	Avaliação ou Modelagem do Estudante	Mineração de texto	Rooda
RABBANY; TAKAFFOLI; ZAÏANE (2011)	Estimativa ou Modelagem de desempenho de estudante	Mineração de Texto	Moodle

RICARTE; JUNIOR (2011)	Modelagem de Grupos ou Aprendizagem Colaborativa	Clusterização	TIDIA-Ae
GOTTARDO; KAESTNER; NORONHA (2012A)	Estimativa ou Modelagem de desempenho de estudante	Classificação	Moodle
GOTTARDO; KAESTNER; NORONHA (2012B)	Estimativa ou Modelagem de desempenho de estudante	Classificação	Moodle
MANHÃES *et. al* (2012)	Detecção ou Previsão de Evasão	Classificação	SIGA
SILVA; MORINO; SATO (2014)	Avaliação ou Modelagem do Estudante	Classificação	ENADE/INEP
GOTTARDO; KAESTNER; NORONHA (2014)	Estimativa ou Modelagem de desempenho de estudante	Classificação	Moodle
KAMPFF *et al*.(2014)	Estimativa ou Modelagem de desempenho de estudante	Regras de Associação	NetAula
SANTANA; MACIEL; RODRIGUES (2014)	Avaliação ou Modelagem do Estudante	Classificação	Moodle

Dringus (2005) buscou novos indicadores de participação a partir da análise de fórum em uma AVA utilizando técnicas de Mineração de Texto para o Inglês. Este trabalho apresenta a possibilidade de aplicação de TM em busca de conhecimento em textos em AVAs, inspirando esta dissertação a incluir análise de características educacionais no estudo de caso proposto.

No campo da análise da natureza social do aprendizado, os trabalhos de Macfadyen; Dawson (2010), Kampff *et al*.(2010), Rabbany, Takaffoli; Zaïane (2011) focam na participação e interação dos estudantes em um AVA, no qual Rabbany *et al* (2011) investiga a importância do estudo de redes sociais, utilizando a técnica de "mineração de comunidades" em busca de estruturas relevantes em fóruns de discussão.

Vale ressaltar que os trabalhos supracitados valorizam a busca por perfil de grupos de alunos de maneira a caracterizar e identificar alunos com perfis semelhantes de maneira a propiciar metodologias de ensino personalizadas. Desta forma, o estudo de caso utilizado nesta dissertação também busca encontrar padrões de perfis de alunos, porém diferente dos

trabalhos supracitados a análise é feita com base em agrupamento nos logs de interação dos alunos, tal como o trabalho de Ricarte (2011) que utilizou clusterização com os algoritmos K-*means* e Mapas Auto-Organizáveis (KOHONEN, 1995) em uma base de *logs* coletada a partir do uso de um AVA na UNICAMP. O autor almejou encontrar grupos de estudantes com comportamento semelhante; oferecendo retorno a autores e tutores sobre o uso dos conteúdos disponibilizados, e prover aos estudantes as informações sobre seu próprio uso dos recursos do ambiente.

Na área de detecção de evasão, com o objetivo de identificar precocemente o subconjunto de alunos que apresentam risco de evasão, Manhães *et. al.*(2011) aplicaram técnicas de Mineração de Dados em uma base de dados coletada do sistema acadêmico da UFRJ, contendo as informações acadêmicas referentes ao primeiro período letivo de alunos de Engenharia da UFRJ de curso presencial. Foram aplicados dez algoritmos de classificação da ferramenta Weka, a fim de avaliar o desempenho destes ao domínio supracitado. Porém tal qual os demais correlatos supracitados nessa dissertação, o tranalho de Manhães *et al.* (2011) não possui metodologia de aplicação das técnicas.

Dos trabalhos correlatos apresentados, apenas o de trabalho de Beer; Clark; Jones (2010) se diferenciava da abordagem de aprendizado de máquina, desenvolvendo análise estatística dos dados de notas dos alunos, média de acesso por página, média de tempo de permanência nos AVAs na apresentação de indicadores de engajamento. Apesar da abordagem estatística, os autores também não apresentam metodologia consolidada.

Dado o levantamento preliminar percebeu-se a necessidade de seleção de atributos que representem bem o objetivo educacional. Assim como Gottardo, Kaestner e Noronha (2012a) e Santana, Maciel e Rodrigues (2014), esta dissertação baseia-se nas indicações de Moore (1989), especificamente na dimensão de "perfil de uso do AVA" para a escolha dos atributos a serem utilizados na metodologia aqui proposta. E do mesmo modo que Romero, Ventura e García (2008), a metodologia proposta nesta dissertação cria uma tabela resumo contendo atributos genéricos dos estudantes e professores.

Em relação à avaliação de classificadores em tarefas de EDM, notou-se a que a acurácia − com destaque para o método *K-cross validation* − e a matriz de confusão estão consolidadas como medida de desempenho dos classificadores (GOTTARDO; KAESTNER; NORONHA, 2012A) (MACIEL; RODRIGUES, 2014) (ROMERO; VENTURA; GARCÍA, 2008).

Por meio da análise dos trabalhos acima apresentados percebe-se que a falta de padronização de dados, além de prejudicar uma modelagem geral e comparação dos trabalhos,

dificulta a adoção destas metodologias de extração de conhecimento de dados educacionais em ambientes reais para contribuição no processo de ensino aprendizagem.

4.4 CONSIDERAÇÕES FINAIS

Neste capítulo apresentaram-se os principais trabalhos que abalizaram a proposta desta dissertação. Inicialmente, discutiram-se os direcionamentos das pesquisas envolvendo Mineração de Dados Educacionais para então expor as principais abordagens desenvolvidas. Neste ensejo, percebeu-se a carência por uma metodologia genérica para EDM e até mesmo, a aplicação de métodos de EDM em larga escala propiciando maior conhecimento do processo de aprendizagem. Foi possível também, por meio da análise comparativa entre os métodos experimentais destes trabalhos, identificar a falta de padronização, abrangendo: conjuntos de dados utilizados, medidas de avaliação das técnicas desenvolvidas, e ainda, nos testes estatísticos para avaliação dos modelos. Este cenário torna notória a necessidade de desenvolvimento de uma proposta para padronização dos experimentos envolvendo EDM, de forma a propiciar a comparação fidedigna entre as pesquisas desenvolvidas na área.

Estas lacunas identificadas na literatura serviram de subsídio para o desenvolvimento da presente dissertação, a qual é apresentada com maiores detalhes no próximo capítulo.

5 METODOLOGIA DE AVALIAÇÃO DE DESEMPENHO EM PROGRAMA DE FORMAÇÃO MASSIVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

5.1 CONSIDERAÇÕES INICIAIS

Como evidenciado nos capítulos anteriores, a problemática de falta de padronização de métodos e dados educacionais afeta negativamente aplicações de metodologias genéricas em diferentes contextos. A falta de padronização nos experimentos envolvendo sistemas educacionais dificulta a comparação entre os diversos métodos propostos e, também, a adoção dos novos métodos neste domínio.

Baseado nas hipóteses de que: i) a adoção de uma padronização nos experimentos envolvendo dados educacionais aumenta a qualidade das pesquisas na área, impulsionando/propiciando a plena utilização dos métodos recentemente propostos; ii) e que o contexto do processo de aprendizagem pode influenciar no desempenho dos estudantes; desenvolveu-se a pesquisa que será apresentada a seguir.

Este capítulo também aponta algumas contribuições com o desenvolvimento do trabalho, bem como apresentação de estudo de caso de aplicação da metodologia proposta com dados reais.

5.2 METODOLOGIA

A pesquisa realizada divide-se na metodologia e estudo de caso, como apresentado a seguir:

5.2.1 **Metodologia para Mineração de Dados Educacionais**

Como exposto no capítulo anterior, há uma lacuna na literatura quanto à padronização de experimentos de extração de conhecimento de dados educacionais. Em um primeiro

momento, buscou-se estabelecer os passos do processo de padronização de Mineração de Dados Educacionais a fim de introduzir informações do contexto educacional e guiar a escolha de atributos a serem utilizados no processo.

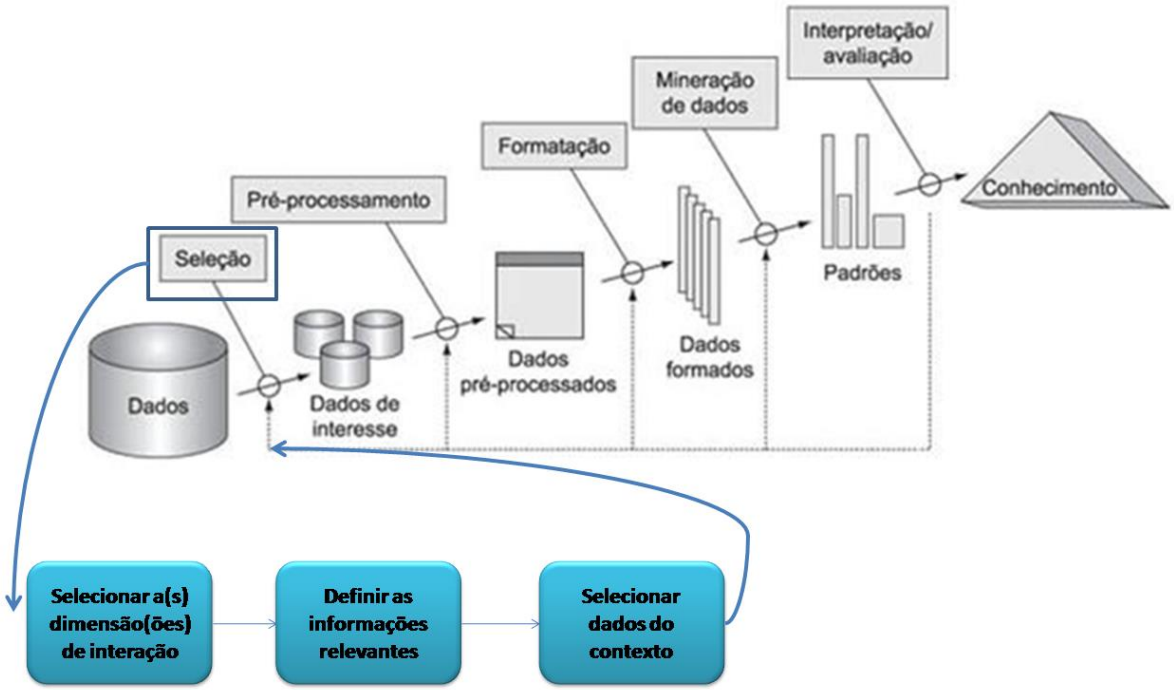

Figura 5.1 – Metodologia de Mineração de Dados Educacionais proposta.

Pensou-se o modelo das tarefas da metodologia como cascata, como mostra a Figura 5.1 já que as atividades influenciam as subsequentes. O processo da metodologia de padronização de Mineração de Dados Educacionais é um processo de especialização do processo de KDD, com ênfase e adição de tarefas na etapa de seleção de atributos considerando o contexto educacional e dimensão de interação.

O processo inicia tal qual o processo de KDD com a aquisição e conhecimento dos dados. Há o levantamento de hipóteses e possíveis perguntas a serem, como por exemplo, qual o perfil de alunos? Qual o conteúdo de interesse dos alunos? Qual as dificuldades do aluno? É possível predizer o desempenho do aluno? É possível predizer evasão? Etc.

Conseguinte, há a aplicação da metodologia proposta nesta dissertação como tarefa de seleção de atributos para dados educacionais, a qual consiste: i) definição de qual tipo de

dimensão e informações deseja-se extrair conhecimento com base na Teoria da Interação; ii) definição de quais informações são importantes considerando o contexto educacional; iii) seleção de atributos genéricos que representem as informações da etapa anterior. De forma a melhorar os resultados levando em consideração o contexto, deve-se selecionar os dados que representem o contexto educacional.

A metodologia proposta difere do processo padrão de seleção de dados do KDD pois é específica para dados educacionais pois tendência a escolha dos atributos para dimensões de interação e dados que caracterizem o contexto de ensino (perfil, avaliações, dados socioeconômicos, mensagens, etc.) dentro do ambiente de aprendizagem, enquanto que o processo padrão de KDD seleciona os melhores atributos que caracterizam determinado problema, dependente ou não de contexto.

Conseguinte, deve-se realizar o pré-processamento, limpeza, transformações e escolher e aplicar técnicas de Mineração de Dados. A última tarefa do processo é a avaliação e interpretação dos resultados para extração de conhecimento dos dados educacionais levando em consideração o contexto. Nesta fase é importante considerar quais métricas de avaliação serão utilizadas para correlação aos fatores que influenciam os resultados. É importante destacar que nos trabalhos correlatos foram apresentadas algumas métricas usuais e potencialmente úteis para esta dissertação.

A motivação de proposta de metodologia dá-se em conta à falta de padronização de Mineração de Dados Educacionais como é reportado pela literatura. Visando avaliar esta hipótese, aplica-se a metodologia apresentada nesta dissertação com estudo de caso de Avaliação de Desempenho de alunos na Formação do Telecentros.BR utilizando Mineração de Dados nas bases de dados do Moodle e Sistema de Avaliação do programa.

5.2.2 **Avaliação de Desempenho em Programa de Formação Massiva utilizando Mineração de Dados**

Como exposto nos trabalhos correlatos, há uma série de trabalhos que realizam a avaliação ou estimativa de desempenho de estudantes, porém não há padronização ou correlação de contexto educacional nos referidos estudos.

A presente dissertação tem por objetivo proposta de metodologia de padronização de seleção de atributos para EDM. Para verificação da viabilidade da metodologia proposta, utilizou-se de estudo de caso para avaliação de desempenho em programa de Formação Massiva utilizando metodologia de Mineração de Dados Educacionais. Neste ensejo, pensou-se na metodologia do estudo apresentada na seção anterior.

Para estudo de caso, os dados utilizados no presente trabalho foram coletados durante o processo de Formação do Programa Telecentros.BR através do Moodle e Sistema de Avaliação de Monitores. Para comprovação da hipótese de que o contexto do processo de aprendizagem pode influenciar no desempenho dos estudantes, este estudo utilizou as bases de dados do Moodle e a base do Sistema de Avaliação de Monitores utilizados no Telecentros.BR para que dados provenientes da Formação e aspectos do contexto educacional pudessem ser correlacionados.

O primeiro passo foi compreender o processo de Formação e qual o tipo de dados armazenados. A base de dados do Moodle utilizado na Rede Telecentros BR por aproximadamente mil e quatrocentos alunos (monitores) armazenou todas as interações dos usuários na plataforma em forma de *logs*. A coleta de dados ocorreu de maneira implícita através do uso e interação dos monitores na plataforma. Uma vantagem de registrar as atividades de um usuário em uma plataforma em forma de log é que grande grandes volumes de dados podem ser armazenados automaticamente (ROGERS; SHARP; PREECE, 2011). Os *logs* armazenados pelo Moodle apresentam características da Web: URLs, texto, números, mídias, etc.

O Sistema de Avaliação da Formação de Monitores apresenta uma base de dados relacional com 38 tabelas, onde são armazenados os dados em relação aos tutores e seu grupo de monitores e as avaliações por cada módulo realizado pelos monitores. A base apresenta 12598 avaliações feitas para aproximadamente mil e quatrocentos monitores ao longo da formação. Destas avaliações, 7612 foram avaliações de 44% de monitores que não participaram dos módulos avaliados e 4986 avaliações de 66% de monitores que participaram totalmente ou parcialmente dos módulos. A base contém 62642 respostas objetivas e 53550 observações em relação à participação dos monitores nos tópicos avaliados. Os dados apresentados nesta base são discretos e textuais.

Após o entendimento do domínio de aplicação foram definidas as perguntas as quais desejou-se responder sobre o processo de formação do Telecentros.Br. São elas:

1. Encontrar perfis de alunos a partir dos logs de uso do Moodle, com destaque para os recursos mais utilizados na plataforma utilizando o algoritmo para agrupamento *K-Means* com utilização de técnicas de DM *Web Mining*;
2. Fazer levantamento estatístico dos desempenhos dos alunos na formação através dos conceitos no Sistema de Avaliação;
3. Encontrar características educacionais a partir das observações qualitativas do Sistema de Avaliação utilizando agrupamento através do algoritmo SOM com utilização de técnicas de *Text Mining*.
4. Identificar relações entre os perfis de uso, desempenho e características educacionais dos alunos;
5. Testar classificadores baseados em Árvore de Decisão na base de avaliação para elencar potenciais métodos para predição de desempenho de alunos no processo de Formação

Passando para a etapa de aplicação da metodologia, definiu-se primeiramente a dimensão de interação dos dados relevantes para esta pesquisa como perfil de uso do AVA, a fim de se obter conhecimento a partir do que comportamento durante a formação e correlacionar com o contexto educacional. O segundo passo foi definir os dados relevantes para o cenário educativo analisado como informações de identificação de usuário, curso, módulo, ações e notas das avaliações realizadas no processo de aprendizagem.

Partindo para a terceira etapa da metodologia, foram definidos os atributos para representação das informações do passo anterior. Para tal, definiram-se os atributos da tabela mdl_log para representação das informações citadas, como mostra a Tabela 5.1, atendendo assim o terceiro passo da metodologia proposta.

Tabela 5.1 – Atributos utilizados padronizar

Atributo	Descrição	Tipo
Userid	Identificador do usuário	Discreto
Course	Identificador do curso	Discreto
Module	Módulo Acessado	Nominal

Action	Ação realizada	Nominal

Com base em análise prévia da base de dados do Sistema de Avaliação, para as informações de contexto foram utilizadas os dados deste sistema, os quais contêm informações qualitativas a respeito do processo de formação, caracterizando o contexto educacional. Então foram selecionadas as seguintes tabelas da base de dados mostradas na Tabela 5.2. Os dados contidos nas tabelas *tt_resposta_opcao* e *tt_resposta* apresentam dados numéricos discretos e dados textuais, respectivamente, representando informações da utilização de recursos tecnológicos, participação em redes sociais, movimento nos temas, colaboração e processo formativo.

Tabela 5.2 – Tabelas selecionadas da base de avaliação

Tabela do Banco de Dados	Descrição
tt_avaliacao	Armazena os identificadores das avaliações, tutor e monitor, data da avaliação, ausência de monitor e motivo de ausência
tt_resposta_opcao	Armazena as respostas das perguntas objetivas fechadas das avaliações tendo como respostas fechadas: Apresenta facilidade, não apresenta facilidade, não, não consigo avaliar
tt_resposta	Armazena as respostas das perguntas subjetivas abertas das avaliações tendo como respostas texto expressando o parecer do tutor em relação ao monitor utilizando Linguagem Natural
tt_usuario	Armazena as informações sobre os usuários do sistema

A partir desse entendimento inicial do processo de Formação do Telecentros.BR e questões a serem respondidas pelo processo de EDM, esta pesquisa iniciou os experimentos em busca dos registros de uso de recursos da plataforma, denominados de caminhos médios, mais utilizados pelos monitores na formação (Pinheiro, et al., 2014a) a fim de encontrar os padrões de uso dos recursos mais utilizados por grupos de monitores encontrados correlacionando com o contexto educacional, como descrito em Pinheiro et al. (2014b).

Para a realização das tarefas citadas acima, optou-se por manter o sigilo e preservar a identidade dos alunos, então foi criada uma tabela "resumo" contendo os dados de acesso, dados geográficos e da formação em uma única tabela, excluindo-se o atributo "*userid*" e quaisquer informações pessoais do aluno, foi acrescentado o campo "*id*" identificando a instância, este campo foi gerado de forma incremental e supervisionada. Dessa forma, não foram considerados dados pessoais dos alunos, somente o perfil de uso e dados geográficos deste. A Tabela 5.3 apresenta os dados resumidos em tabela única

Tabela 5.3 – Resumo Perfil de Acesso.

Fonte	Atributo	Descrição	Tipo
	Id	Identificador de perfil	Discreto
Moodle	Course	Identificador do curso	Discreto
Moodle	Module	Módulo Acessado	Nominal
Moodle	Action	Ação realizada	Nominal
Sistema de Avaliação	UF	Estado do aluno	Nominal
Sistema de Avaliação	Cidade	Cidade do Aluno	Nominal

Da mesma forma, foi criada uma tabela resumo para os dados do Sistema de Avaliação dos Monitores, apresentada na Tabela 5.4, de maneira a concentrar a informação e criar um perfil educacional do aluno. Vale ressaltar a relação do perfil de acesso no Moodle com o perfil educacional através do atributo *id*, com relação de 1 para *n*, onde cada *id* pode estar ligado a *n* avaliações.

Tabela 5.4 – Perfil Educacional

	Id	**Identificador de perfil**	**Discreto**
Sistema de Avaliação	Uso do Moodle	Avaliação de uso do Moodle	Nominal
Sistema de Avaliação	Uso de Software Livre	Avaliação de uso de Software	Nominal
Sistema de Avaliação	Tecnologias Disponíveis	Avaliação de Tecnologias Disponíveis	Nominal
Sistema de Avaliação	Acesso à Internet	Avaliação de Uso da Internet	Nominal
Sistema de Avaliação	Presença em redes sociais	Avaliação de presença em redes sociais	Nominal
Sistema de Avaliação	Articulação Interativa	Avaliação de interação nas redes sociais	Nominal
Sistema de Avaliação	Colaboração	Avaliação de colaboração	Nominal
Sistema de Avaliação	Projeto Comunitário	Avaliação da participação em projeto	Nominal
Sistema de Avaliação	Aprendizado de conteúdo	Avaliação do aluno (Classe)	Nominal
Sistema de Avaliação	Utilização de Recursos Tecnológicos	Observação sobre utilização de Recursos Tecnológicos	Textual

Sistema de Avaliação	Participação em Redes Sociais	Observação sobre interação nas redes sociais	Textual
Sistema de Avaliação	Movimento nas Zonas Temáticas	Observação sobre colaboração	Textual
Sistema de Avaliação	Colaboração	Observação sobre colaboração	Textual
Sistema de Avaliação	Processo Formativo	Observação sobre processo Formativo	Textual

Após a definição dos atributos. Partimos para a etapa de Mineração de Dados, particionada de acordo com os objetivos requeridos.

5.2.3 **Perfis de uso**

Para alcançar a primeira tarefa definida como 1) Encontrar perfis de alunos a partir dos logs de uso do Moodle, com destaque para os recursos mais utilizados na plataforma utilizando o algoritmo para agrupamento *K-Means* com utilização de técnicas de DM *Web Mining*, foram utilizados os dados apresentados na Tabela 5.3. O pré-processamento consistiu na limpeza das informações das demais tabelas do Moodle em conjunto com o filtro para eliminar alunos excluídos e administradores do sistema procurando sanar a interferência destes grupos de usuários.

A partir da etapa de pré-processamento a clusterização é permitida usando o algoritmo *K-Means* implementado em SQL como rotina diretamente no Sistema Gerenciador de Banco de Dados (SGBD) MySQL para duas dimensões onde são tratadas as colunas: "*course*" e "*id*", contando o número de acessos de cada usuário para cada curso, também são filtrados somente eventos dentro da categoria "*course*".

Após a composição do *dataset*, o algoritmo *K-Means* fora executado para k=5, pois são estabelecidos pelo programa quatro níveis de avaliações qualitativas: Excelente, Bom, Regular e Insuficiente, contudo foi acrescentada uma classe para que *outliers* possam ser minimizados nas outras classes (FAYYAD et al., 1996).

O algoritmo foi executado em 13283 linhas de *log* do Moodle e obtiveram-se cinco clusters, como é mostrado a seguir na Tabela 5.5.

Tabela 5.5 – *Clusters* encontrados por perfil de uso.

	K1	**K2**	**K3**	**K4**	**K5**
Quantidade Usuários	5	32	183	825	3251
Número de Acessos	228728	598867	1197017	2024758	1152300
Cursos	50	50	50	49	50
Média de acesso por usuário	45745,6	18714,5	6541,1	2454,2	354,4
Média de acesso por usuário e por curso	914,9	374,2	130,8	50	7
Média de acesso por usuário e por curso	View 93930	View 266186	View 614599	View 1183165	View 93930
Recursos mais acessados	Error Login 59430	Course 153197	Forum 294337	Course 513217	Course 310376
Caminhos médios	145	163	167	137	133

5.2.4 **Desempenho dos alunos**

A fim de sanar a segunda tarefa definida no processo de EDM, 2) Fazer levantamento estatístico dos desempenhos dos alunos na formação através dos conceitos no Sistema de Avaliação, em busca de aspectos que estivessem além das características da avaliação proposta

pelos gestores da Rede de Formação Telecentros.Br, foram definidos como conceitos finais dos alunos o atributo "*AprendizadodeConteudo*".

Os valores dos conceitos dados aos alunos originalmente eram definidos por "*Não Consigo Avaliar*", "*O monitor ainda revela um entendimento superficial dos conteúdos trabalhados*", "*O monitor revela razoável entendimento dos conteúdos explorados*" e "*O monitor revela um bom entendimento acerca dos conteúdos*" sendo estes transformados de forma supervisionada, respectivamente, para as classes "*Insuficiente*", "*Regular*", "*Bom*" e "*Excelente*".

A partir de análise estatística dos dados, verificaram-se os seguintes dados quantitativos em relação ao conceito final dos alunos, mostrados na Tabela 5.6.

Tabela 5.6 – Conceitos definidos para a classificação

Conceito	Quantidade
Insuficiente	1312
Bom	1123
Regular	769
Excelente	1761

Nota-se que há uma alta taxa de alunos que apresentaram desempenho "*Insuficiente*" e de alunos que apresentaram "*Bom*" ou "*Excelente*" na formação. Visando encontrar as causas para estas taxas de conceitos então seguiu-se para a tarefa de encontrar caracteríscas educacionais que influenciaram no processo de aprendizagem.

5.2.5 **Encontrar características educacionais**

Devido à alta taxa de desempenho de alunos com conceito "Insuficiente", "*Bom*" ou "*Excelente*" , buscou-se encontrar indicadores para os conceitso através da análise de avaliação qualitativa, chamadas de observações textuais do processo de Formação a partir dos atributos "Utilização de Recursos Tecnológicos", "Participação em Redes Sociais", "Movimento nas Zonas Temáticas", "Colaboração" e "Processo Formativo" informados na Tabela 5.4 com intuito de completar a tarefa 3) Encontrar características educacionais a partir das observações qualitativas do Sistema de Avaliação utilizando agrupamento através do algoritmo SOM com utilização de técnicas de *Text Mining*.

As observações textuais no Sistema de Avaliação eram observações realizadas pelos tutores do Telecentros.br com o intuito de acrescentar mais informações a respeito do processo de formação individual dos monitores. Essas observações foram todas escritas em linguagem natural utilizando o Português Brasileiro (PT-Br) e totalizaram 68174 observações realizadas durante o processo de Formação. A Figura 5.2 ilustra uma amostra das observações realizadas.

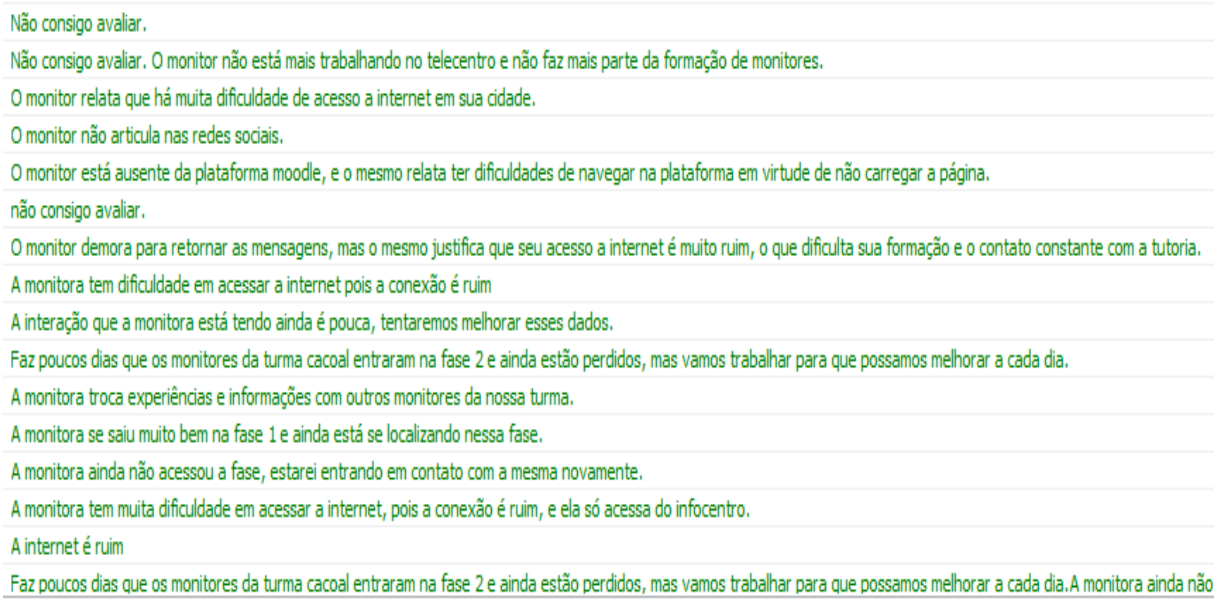
Não consigo avaliar.
Não consigo avaliar. O monitor não está mais trabalhando no telecentro e não faz mais parte da formação de monitores.
O monitor relata que há muita dificuldade de acesso a internet em sua cidade.
O monitor não articula nas redes sociais.
O monitor está ausente da plataforma moodle, e o mesmo relata ter dificuldades de navegar na plataforma em virtude de não carregar a página.
não consigo avaliar.
O monitor demora para retornar as mensagens, mas o mesmo justifica que seu acesso a internet é muito ruim, o que dificulta sua formação e o contato constante com a tutoria.
A monitora tem dificuldade em acessar a internet pois a conexão é ruim
A interação que a monitora está tendo ainda é pouca, tentaremos melhorar esses dados.
Faz poucos dias que os monitores da turma cacoal entraram na fase 2 e ainda estão perdidos, mas vamos trabalhar para que possamos melhorar a cada dia.
A monitora troca experiências e informações com outros monitores da nossa turma.
A monitora se saiu muito bem na fase 1 e ainda está se localizando nessa fase.
A monitora ainda não acessou a fase, estarei entrando em contato com a mesma novamente.
A monitora tem muita dificuldade em acessar a internet, pois a conexão é ruim, e ela só acessa do infocentro.
A internet é ruim
Faz poucos dias que os monitores da turma cacoal entraram na fase 2 e ainda estão perdidos, mas vamos trabalhar para que possamos melhorar a cada dia.A monitora ainda não

Figura 5.2 – Amostra da base de observações realizadas.

Para extração de conhecimento dessas observações aplicou-se Mineração de Texto com utilização de SOM com embasamento na revisão de literatura feita e pelo número desconhecido de grupos de comportamento qualitativo a serem encontrados.

Para tal, o pré-processamento de texto consistiu na separação de *tokens*, retirada de pontuação e caracteres especiais, retirada de acentuação, padronização com letras minúsculas, remoção de *stop words* e normalização morfológica. Optou-se por utilizar a ferramenta Apache Lucene[1] pela disponibilidade das tarefas de pré-processamento para Português Brasileiro (PT-Br) e implementação de aplicação de Mineração de Texto em linguagem de programação Java pelo domínio da linguagem. Após as observações passarem por pré-processamento, obteve-se

[1] Apache Lucene: https://lucene.apache.org/core/

como saída uma coleção de documentos formatados e pré-processados, estes foram armazenadas em uma tabela em banco de dados.

Por se tratar de linguagem natural, a alta dimensionalidade dos dados nesta base é evidente, já que a base apresenta 5682 palavras distintas em PT_BR. Para visualizar a dimensão das observações, foi plotado o gráfico de distribuição das observações por questões e avaliações realizadas, exibida na Figura 5.3, onde o gradiente de cor representa os diferentes grupos de respostas.

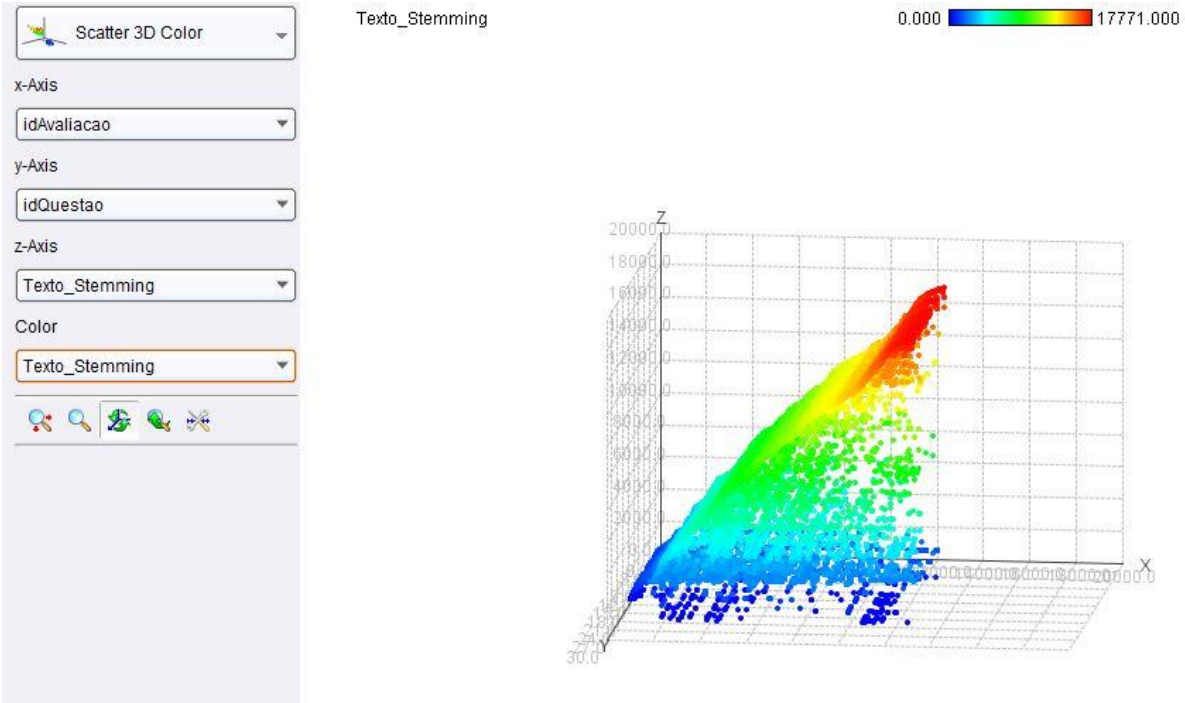

Figura 5.3 – Distribuição das observações por avaliações realizadas.

Dada a alta dimensionalidade dos dados, optou-se por utilizar o modelo de espaço vetorial –VSM (SALTON; MCGILL, 1983) TF*IDF, visando à redução de dimensionalidade e boa representação dos dados.

O primeiro passo da etapa de pré-processamento consistiu na remoção de *stop words*, vale ressaltar que fora retirada a palavra "*não*" da lista de *stop words* padrão do Lucene pois esta poderia representar dificuldades no processo de formação. Em seguida realizou-se o

processo de *stemming*, reduzindo às palavras às suas respectivas raízes morfológicas, como ilustra a Figura 5.4.

nao consig avali monitor nao trabalh no telecentr nao faz part formaca monitor
monitor relat muit dificuldad acess internet cidad
monitor nao articul red soci
monitor ausent plataform moodl relat ter dificuldad naveg plataform virtud nao carreg pagin
nao consig avali
monitor demor par retorn as mensagens justif acess internet muit ruim dificult formaca contat constant tutor
monitor tem dificuldad acess internet conexa ruim
inter monitor tend pouc tent melhor dad
faz pouc dias monitor turm cacoal entrar fas esta perd vam trabalh par poss melhor cad dia

Figura 5.4 – Palavras mais frequentes nas avaliações do Monitores

Após essas etapas, realizou-se a verificação das palavras mais frequentes e com maior relevância para extração de conhecimento das observações. Então, optou-se pela criação de histogramas para descartar as palavras com frequência na base inferior a 200, totalizando 313 palavras restantes para representação das observações. Algumas das palavras mais frequentes são mostradas na nuvem de palavras ilustrada na Figura 5.5.

Figura 5.5 – Palavras mais frequentes nas avaliações do Monitores

Dentre as palavras com maior frequência nas observações, podemos destacar as palavras que são apresentadas na Tabela 5.7, as quais podem ser caracterizadas como palavras-chave no processo de formação.

Tabela 5.7 –Seis palavras mais frequentes na base de avaliações.

Frequência	Palavra
29603	Monitor
17838	Não
11269	Projeto
8018	Atividade
7895	Acesso
7549	Formação

Após a seleção das palavras a serem utilizadas no processo de TM, foram aplicados os passos de transformação VSM das observações utilizando representação TF*IDF, onde observação é representada por um vetor formado pelo conjunto de palavras as quais a formam e cada palavra é representada pelo seu respectivo valor numérico, como ilustra a Figura 5.6.

Figura 5.6 – Amostra de observação com representação TF*IDF.

Para melhor compreensão do contexto educacional por conceitos dos alunos e obtenção dos grupos de observações qualitativas do Sistema de Avaliação, optou-se por dividir a base de observações pelos cinco conceitos apresentados na seção 5.2.4, fazendo-se necessário processamento para cada grupo de observações.

Para cada grupo de observações foi aplicado o algoritmo SOM formado por um grid de 54 por 54 neurônios, totalizando 2916 neurônios.

No total, 45 *clusters* foram encontrados a partir do processamento da base de observações. Os grupos encontrados apresentam grupos de sentenças semelhantes e que permitem analisar as características de cada perfil de alunos durante a Formação.

Buscou-se, principalmente, obter informações das causas do conceito “Insuficiente”. Três grupos de observações para este conceito chamaram atenção devido a fatores socioeconômicos de evasão, como gravidez, emprego e dificuldade de acesso à Internet, como ilustram as Figuras 5.7, 5.8 e 5.9, respectivamente.

Figura 5.7 – Grupo de observações que mostra a gravidez como causa de evasão da Formação.

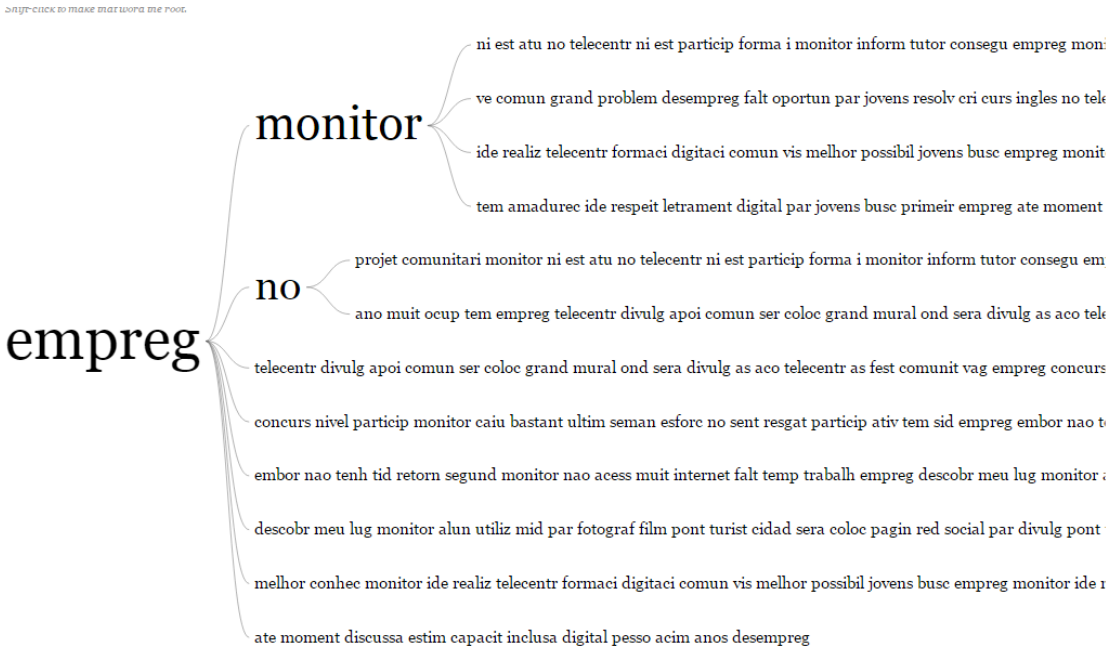

Figura 5.8 – Grupo de observações que mostra o emprego como causa de evasão da Formação.

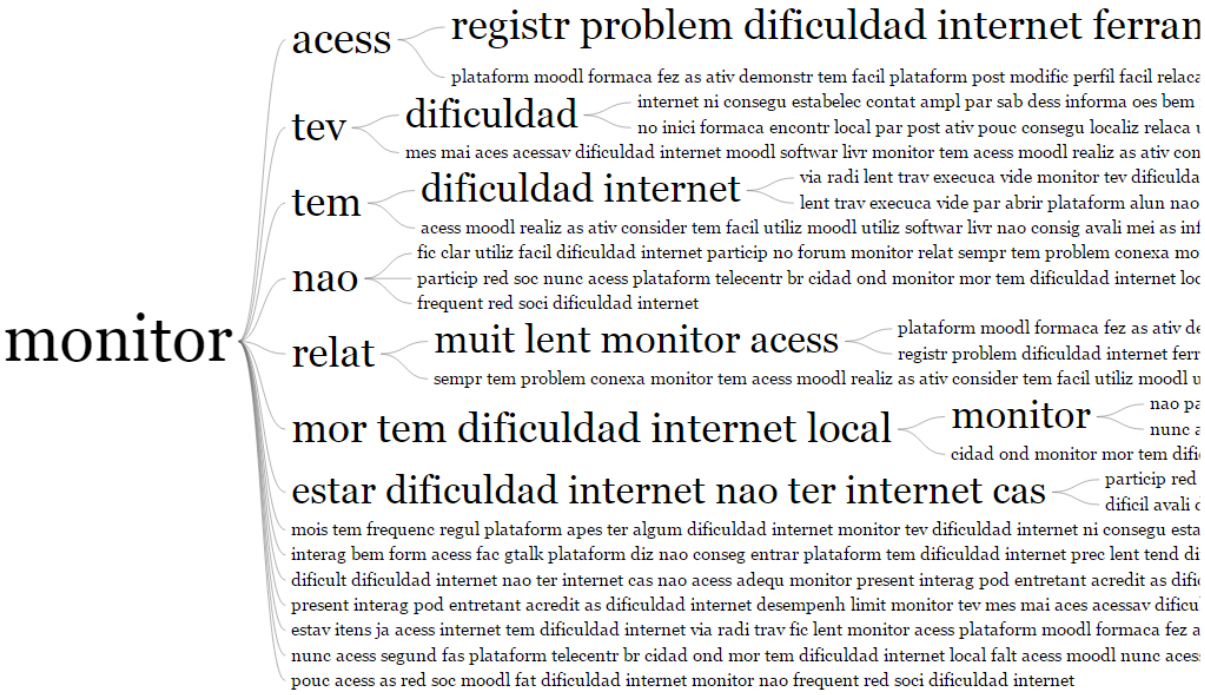

Figura 5.9 – Grupo de observações que mostra dificuldades no acesso à internet como causa de evasão da Formação.

5.2.6 **Testar classificadores**

Posteriormente, com intuito de elencar potenciais classificadores para predição do desempenho de alunos do Telecentros.BR considerando o contexto educacional foi realizada a tarefa 5) Testar classificadores baseados em Árvore de Decisão na base de avaliação para elencar potenciais métodos para predição de desempenho de alunos no processo de Formação.

A tarefa supracitada justifica-se pela possibilidade dos dados nominais de perfil educacional “Uso do Moodle”, “Uso de Software Livre”, “Tecnologias Disponíveis”, “Acesso à Internet”, “Presença em redes sociais”, “Articulação interativa”, “Colaboração”, “Projeto Comunitário” e “Aprendizado de conteúdo”, apresentados na Tabela 5.4, fornecerem possíveis padrões de predição de conceito na Formação. O atributo “Aprendizado de conteúdo” fora escolhido como classe pois no Sistema de Avaliação este atributo representa o real aprendizado do aluno, conforme ilustra a Figura 5.10.

id	UsoMoodle	UsoSoftware	TecnologiasAc	AcessoInternet	freqRedesSociais	publicaAtuacao	articulacaoInter	realidadeMc	processoColabo	Nota	Conceito
1	não consigo aval	com facilida	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
2	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
3	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
4	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
5	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
6	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
7	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
8	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
9	não consigo aval	não consigo	não consigo av	com dificuldade	com frequência	sim	não	Não consigo	não consigo aval	Não consigo aval	Insuficiente
10	com facilidade	com facilida	com facilidade	com dificuldade	com frequência	sim	sim, mas ainda i	Não consigo	o monitor não c	o monitor revela	Bom
11	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	o monitor colab	o monitor ainda	Regular
12	com dificuldade	com dificulc	com dificuldac	com dificuldade	com pouca frequ	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
13	com dificuldade	com facilida	com dificuldac	com dificuldade	não	não	não	Não consigo	o monitor conse	o monitor revela	Bom
14	com facilidade	com facilida	com facilidade	com dificuldade	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
15	não consigo aval	não consigo	não consigo av	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
16	com dificuldade	com dificulc	com dificuldac	com dificuldade	com pouca frequ	não	não	Não consigo	o monitor conse	Não consigo aval	Insuficiente
17	com facilidade	com facilida	com facilidade	com dificuldade	com pouca frequ	não	não	Não consigo	o monitor conse	o monitor revela	Excelente
18	com dificuldade	com dificulc	com dificuldac	com dificuldade	não	não	não	Não consigo	não consigo aval	Não consigo aval	Insuficiente
19	com facilidade	com facilida	com dificuldac	com dificuldade	com pouca frequ	não	não	Não consigo	o monitor colab	Não consigo aval	Insuficiente
20	não consigo aval	não consigo	com facilidade	não consigo ava	não consigo avali	não consigo ava	não consigo ava	Não consigo	não consigo aval	Não consigo aval	Insuficiente
21	com facilidade	com facilida	com facilidade	não consigo ava	não	não	não	Não consigo	não consigo aval	Não consigo aval	Insuficiente
22	com facilidade	com facilida	com facilidade	com dificuldade	com frequência	sim	sim, ele interag	o monitor nã	o monitor acom	o monitor revela	Excelente

Figura 5.10 – Amostra de dados nominais de Perfil Educacional.

Foram escolhidos classificadores baseados em Árvores de decisão para predição de desempenho de alunos do Telecentros.BR pelas vantagens de flexibilidade, capacidade de processamento de valores nominais, robustez, facilidade de interpretação e o modelo "caixa-branca" que permite extração de regras de decisão.

Então, foram utilizados os algoritmos J48, Random Tree da ferramenta Weka[2] nos dados nominais da Tabela 5.4 para testar a potencial utilização de classificadores em ambiente educacional de larga escala. Optou-se pela ferramenta Weka por essa ser uma ferramenta consolidada na área de KDD e por apresentar vários algoritmos de Mineração de Dados do estado da arte desta dissertação. Para a avaliação dos classificadores fora utilizado o método *cross-validation* com 10 *k folds* para avaliação.

Os resultados de acurácias obtidas da tarefa de classificação para os algoritmos propostos são apresentados na Tabela 5.8.

Tabela 5.8 – Avaliação de classificadores em ambiente de Formação Massiva

Algoritmo	Acurácia
J48	72%
Randon Forest	72%

[2] http://www.cs.waikato.ac.nz/ml/weka/

Random Tree	71%

A partir do algoritmo J48 foi possível gerar uma árvore de decisão com regras para predição de desempenho dos alunos do Telecentros.BR, como mostra a Figura 5.11

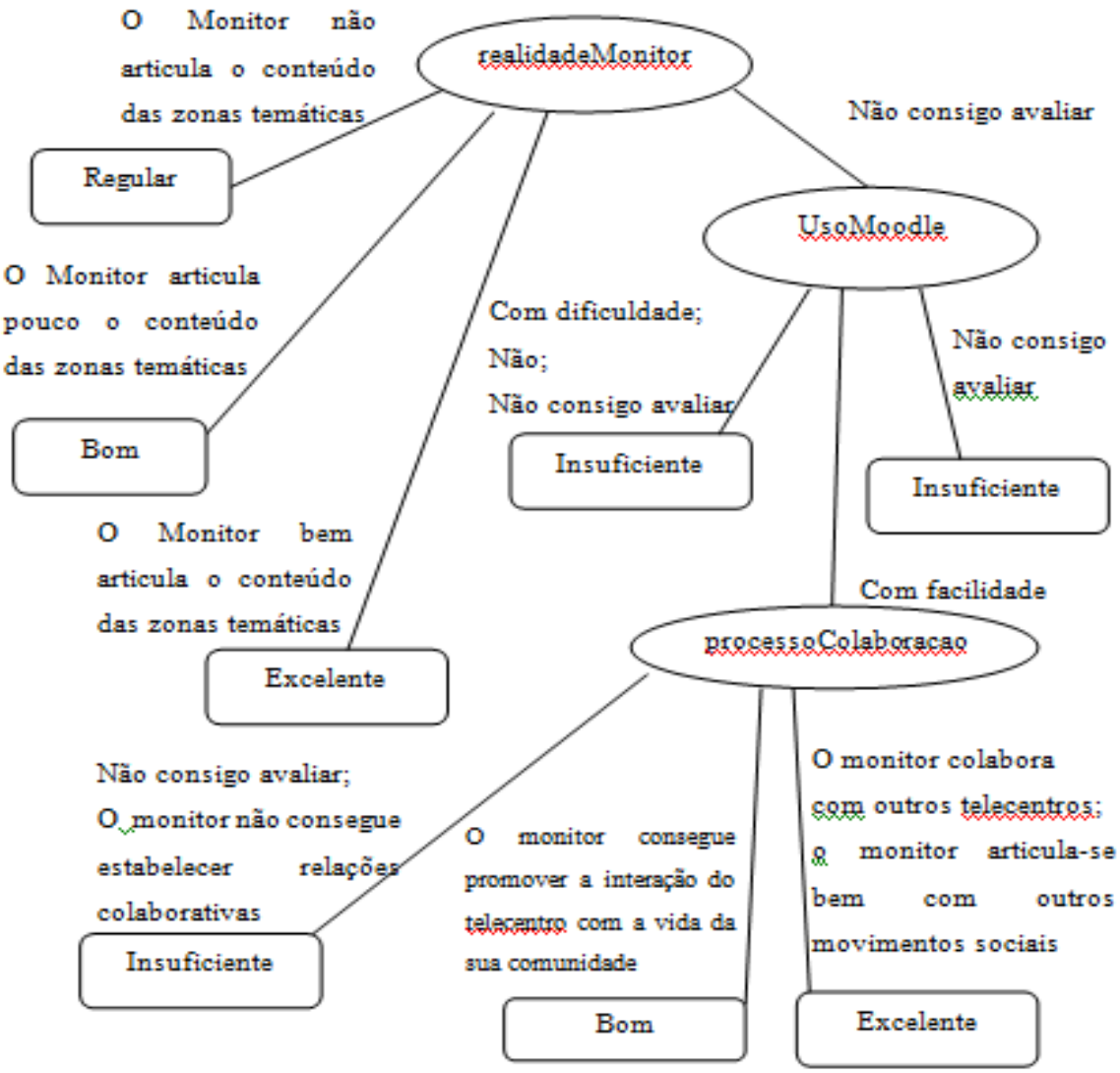

Figura 5.11 – Amostra de dados nominais de Perfil Educacional

5.3 RESULTADOS

O estudo de caso mostrou a viabilidade da utilização da metodologia de Mineração de Dados Educacionais considerando o contexto educacional, sendo importante desde a definição dos atributos até a tarefa de Mineração.

Como prova da eficiência da metodologia, o estudo de caso de Avaliação de Desempenho de Formação Massiva utilizando Mineração de Dados, teve como tarefas:

1. Encontrar perfis de alunos a partir dos logs de uso do Moodle, com destaque para os recursos mais utilizados na plataforma utilizando o algoritmo para agrupamento *K-Means* com utilização de técnicas de DM e Web Mining;
2. Fazer levantamento estatístico dos desempenhos dos alunos na formação através dos conceitos no Sistema de Avaliação;
3. Encontrar características educacionais a partir das observações qualitativas do Sistema de Avaliação utilizando agrupamento através do algoritmo SOM com utilização de técnicas de *Text Mining*.
4. Identificar relações entre os perfis de uso, desempenho e características educacionais dos alunos;
5. Testar classificadores baseados em Árvore de Decisão na base de avaliação para elencar potenciais métodos para predição de desempenho de alunos no processo de Formação

A partir da análise dos resultados das tarefas 1, 2, 3 e 5 é possível realizar a tarefa 4: Identificar as relações entre os perfis de uso; desempenho e características educacionais dos alunos.

Os resultados experimentais da tarefa de clusterizar os *logs* do Moodle do Telecentros.BR mostraram grupos distintos de usuários, sendo possível reconhecer com base nos resultados a distância entre o maior e menor cluster. Isso ocorre, pois no caso do *cluster* K1, pelo pequeno número de usuários e grande número de acessos com erro no Login o grupo demonstra ser formado provavelmente por usuários com comportamento fora do esperado (*outliers*). Buscou-se identificar quais as possíveis causas desse comportamento através da correlação com o desempenho destes alunos, onde se verificou através do desempenho que estes usuários apresentaram o conceito “Insuficiente”, onde não puderam ser avaliados pelos tutores, a partir da Mineração de Texto observou-se que estes monitores evadiram da formação, cujas causas não foram esclarecidas.

No *cluster* K2, identificamos pelo número de usuários e a grande quantidade de acessos, que este cluster é formado pelos usuários responsáveis pela capacitação dos monitores, chamados Tutores. Pudemos comprovar a partir de análise da base de dados de Avaliação que estes usuários eram responsáveis pelas avaliações.

Nos *clusters* K3, K4 e K5 encontramos números de usuários distintos, porém com a média de acesso geral semelhante e com a média por usuário decrescente, respectivamente, então, é possível inferir que estão localizados alunos, chamados monitores, com bom acesso ao sistema no cluster K3, monitores com acesso regular ao sistema no cluster K4 e monitores com um baixo número de acesso na plataforma no cluster K5.

Através de correlação com o desempenho, percebeu-se que os monitores do *cluster* K3, apresentaram avaliação com desempenho "Excelente" ou "Bom" representado por 88% e 12%, respectivamente. Correlacionado com as avaliações qualitativas através de observações textuais e com a árvore gerada pelo algoritmo J48 (Figura 5.8), verificou-se que os alunos são ativos na participação em projetos comunitários, conseguem aplicar os conceitos aprendidos e que são ativos na plataforma Moodle.

Para o *cluster* K4, verificou-se que os monitores apresentam desempenho "Bom" ou "Regular" representado por 54% e 46%, respectivamente. Analisando as avaliações qualitativas desses monitores, percebeu-se que os monitores com desempenho "Bom" são aqueles que conseguem promover a interação do telecentro com a vida da comunidade, estes monitores apresentaram dificuldade de acesso à Internet por ter acesso somente no telecentro. Enquanto os monitores com desempenho "Regular" não conseguiram articular o conteúdo das zonas temáticas.

Para o *cluster* K5 os monitores receberam o conceito "Insuficiente" por falta de contato com os tutores. Investigando-se as causas através das observações, verificou-se que estes monitores já haviam evadido da formação dentre os fatores encontrados temos a ocupação de vaga no mercado de trabalho e a maternidade.

A partir da análise dos recursos mais utilizados, podemos inferir que os monitores do cluster K3 realmente foram mais participativos pelo uso do recurso *Forum*, enquanto que os monitores do cluster K4 tiverem uma maior participação nos materiais dos seus respectivos cursos, assim como os usuários do cluster K5 tiveram participação razoável em seus respectivos cursos. Os caminhos médios foram levantados em busca de encontrar as ações em comum dos usuários de cada cluster na plataforma.

Para os *clusters* que representam os monitores, representados por K3, K4 e K5, obteve-se o uso dos recursos disponíveis para os referidos clusters, como mostram a Figuras 5.12, 5.13 e 5.14, respectivamente.

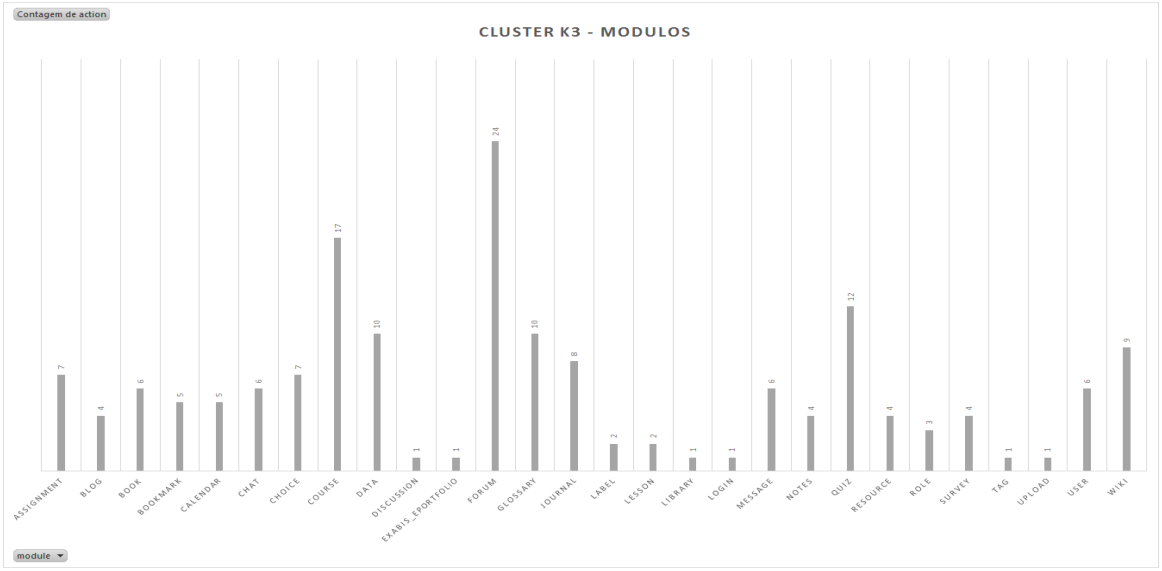

Figura 5.12 – Uso dos recursos pelo cluster K3

Percebe-se que o recurso mais utilizado na plataforma pelos alunos com acesso regular e com baixo acesso, clusters K4 e K5, durante a formação foi o “Curso”, seguido pelo “Recursos” (Arquivos, Mídias, etc.). Enquanto que o recurso de “Fórum” foi o mais utilizado pelos alunos com maior número de acessos à plataforma, seguido pelo “Curso”. Ferramentas importantes de comunicação como Chat, por exemplo, foram muito pouco utilizadas por estes alunos.

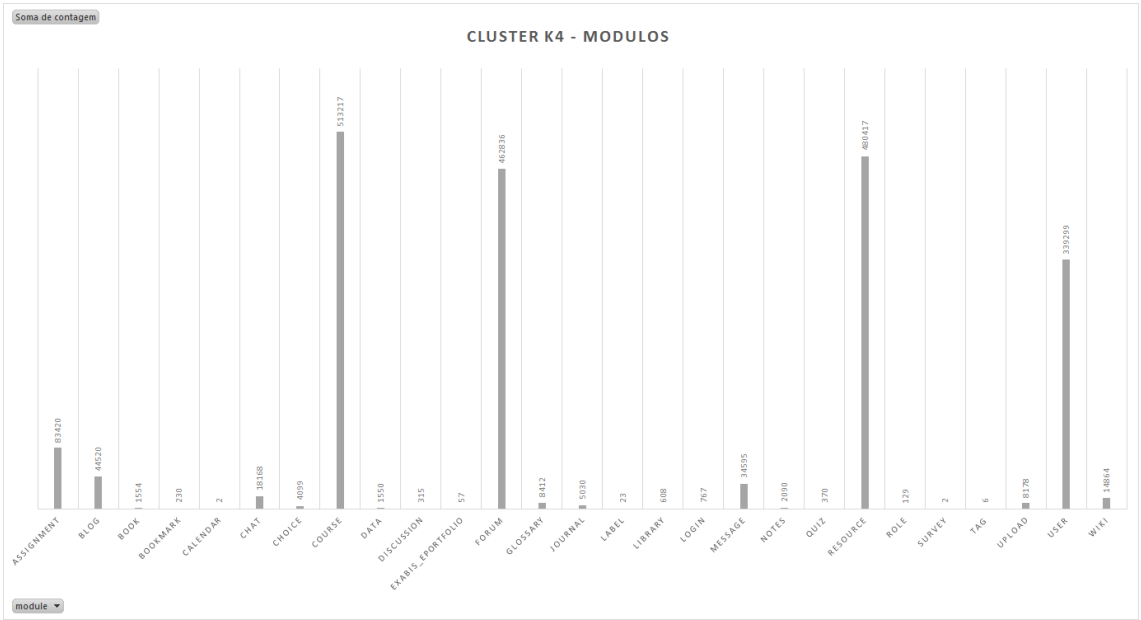

Figura 5.13 – Uso dos recursos pelo cluster K4

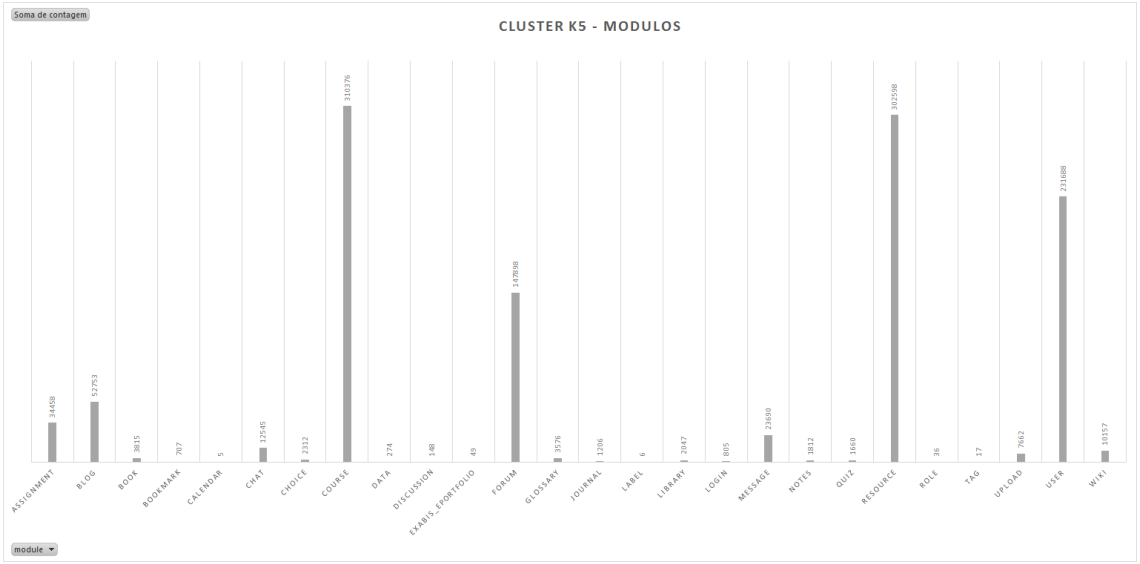

Figura 5.14 – Uso dos recursos pelo *cluster* K5

A partir destes resultados, verificamos a complexidade de Formação Massiva a Distância, já que fatores sociais como desemprego e falta de acesso à Internet podem influenciar na continuação do aluno no programa. Os dados do PNAD (2013) justificam essa realidade, com destaque para a região Norte onde o acesso ocorre principalmente por Internet Móvel.

A tarefa de previsão de desempenho através da classificação mostrou-se viável ao programa, isto porque, os classificadores apresentaram desempenho médio de 72% para os dados utilizados, resultado aceitável para os resultados encontrados em trabalhos correlatos por Hämäläen e Vinni (2011). Além de poder gerar regras para predição de desempenho em formações futuras, como mostrado na Figura 5.11.

Estes resultados mostram que é possível utilizar técnicas de avaliação automática em programas de Formação Massiva de forma a identificar alunos com possível perfil de evasão, ou com dificuldades ao longo do processo, de maneira que gestores e educadores possam ter mais indicadores a respeito do processo de formação e assim, possam tomar decisões em relação às metodologias utilizadas. Assim é possível utilizar técnicas de personalização de ensino com o intuito de alavancar a aprendizagem.

A metodologia proposta mostrou-se viável já que leva em consideração os perfis de interação e contexto educacional propiciando a descoberta de novas relações e indicadores em Formação Massiva, podendo ser generalizada a qualquer AVA ou sistema educacional já que é independente de plataforma e considera o contexto educacional.

Vale ressaltar a importância do uso de avaliação qualitativa no processo de formação, tal como realizado no Telecentros.BR, pois com a visão do tutor em relação ao processo de aprendizagem do aluno é possível identificar o perfil deste último, propiciando assim maior informação sobre o processo educacional.

6 CONSIDERAÇÕES FINAIS

Uma preocupação cada vez mais presente no campo da educação a distância é como melhorar o processo de aprendizagem e diminuir a taxa de evasão. Problema que pode ter solução através da extração de conhecimento da base de dados de sistemas educacionais.

Esta monografia apresentou a dissertação para avaliação da banca examinadora, primeiramente definindo o problema a ser investigado bem como suas motivações, para então definir as hipóteses de trabalho. A primeira versa que a adoção de uma metodologia nos experimentos envolvendo Mineração de Dados Educacionais; enquanto a segunda apresenta novos indicadores aos gestores do Programa Telecentros.BR.

Foram apresentadas as etapas do processo de Descoberta de Conhecimento em Base de Dados, com ênfase nas aplicações em dados Web, dados textuais e educacionais.

Foram discutidas também as lacunas das pesquisas de Mineração de Dados Educacionais, com ênfase na falta de metodologias que padronizem o processo e que promovam uma generalização das aplicações de EDM, proporcionando avanço nas pesquisas através da comparação de resultados e que permitam a criação de modelos genéricos de aprendizagem.

A metodologia proposta neste trabalho considera o potencial da padronização de EDM e o contexto educacional. Como estudo de caso, utilizou-se as bases de dados da Formação Massiva do Telecentros.BR, permitindo que um amplo estudo utilizando técnicas de Mineração de Dados pudesse ser realizado em dados reais. Destaca-se o uso de clusterização para encontrar perfis de uso em AVAs, clusterização para encontrar características qualitativas do processo de formação e classificação a fim de proporcionar a predição de desempenho na Formação Massiva.

As dificuldades encontradas nesta pesquisa estão relacionadas inicialmente ao entendimento do domínio, pois o processo de MineraÇão de Dados Educacionais não é trivial. Seguido das dificuldades operacionais na implementação dos algoritmos de clusterização *K-Means* em SQL e Kohonen em Java. Posteriormente, a tarefa de Mineração de Texto para PT-Br foi de grande desafio, pois são escassas as ferramentas e algoritmos para o idioma, nesta etapa foi necessária a implementação de ferramenta de Mineração de Texto em Java utilizando o pré-processador Apache Lucene.

6.1 CONTRIBUIÇÕES

Esta dissertação gerou como contribuições:

- Revisão de literatura sobre Mineração de Dados Educacionais;
- Metodologia de seleção de atributos de Mineração de Dados Educacionais, tendo como objetivo propiciar a padronização do processo de escolha de atributos que caracterizem o contexto educacional em EDM, possibilitando a aceleração das pesquisas relacionadas e resultados encontrados;
- Estudo de caso promovendo novas formas de avaliação automática do Programa Telecentros.BR utilizando Inteligência Computacional como metodologia para identificar grupos de alunos de maneira a proporcionar uma formação adaptativa por perfil; indicadores socioeconômicos que influenciam no índice de evasão do desempenho de alunos; e metodologia de predição de desempenho de alunos que possibilite o suporte à tomada de decisão quanto às metodologias de ensino aplicadas no Programa Telecentros.BR.

6.2 PUBLICAÇÕES GERADAS

- Pinheiro, M. F., Neto, L. C. F., de Sá Junior, H. N., da Mata, E. C., Jacob Jr, A. F., & de Lima Santana, Á. (2014a). Identificação de Grupos de Alunos em Ambiente Virtual de Aprendizagem Utilizando Análise de Log Baseada em Clusterização. In Anais do XLII Congresso Brasileiro de Educação em Engenharia.
- Pinheiro, M. F., Neto, L. C. F., de Sá Junior, H. N., da Mata, E. C., Jacob Jr, A. F., & de Lima Santana, Á. (2014b). Identificação de Grupos de Alunos em Ambiente Virtual de Aprendizagem: Uma Estratégia de Análise de Log Baseada em Clusterização. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação (Vol. 3, No. 1).

6.3 TRABALHOS FUTUROS

Como trabalhos futuros esta pesquisa almeja:

- A aplicação em outras bases de dados educacionais para comparação de resultados;
- A utilização de técnicas de análise de sequência para encontrar os padrões sequenciais de uso do AVA;
- Avaliar outros classificadores potenciais para a tarefa de predição de desempenho nas bases do Telecentros.BR
- Aplicar regras de associação para descoberta de novos indicadores;
- Utilização de algoritmos voltados para Português Brasileiro para as tarefas de Mineração de texto
- Desenvolvimento de ferramenta de Mineração de Texto para o Moodle
- Desenvolvimento de ferramenta de predição de desempenho de alunos

REFERÊNCIAS

AGGARWAL, Charu C.; ZHAI, ChengXiang. **Mining text data**. Springer Science & Business Media, 2012.

BAEZA-YATES, R.; RIBEIRO-NETO, B. Modern information retrieval. Vol. 463. New York: ACM press, 1999.

BAKER, R.S.J.D. Data Mining for Education. In: McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education. Oxford, UK: Elsevier, 3ed., 2010.

BAKER, R.S.J.D., ISOTANI, S. AND, DE CARVALHO, A.M.J.B. **Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação, vol. 19, no. 2, p. 2-13, 2011.

BECK, J.; WOOLF, B. **High-level student modeling with machine learning**. In: Intelligent Tutoring Systems, pp. 584-593, 2000.

BEER, C.; CLARK, K.; JONES, D. **Indicators of engagement. Curriculum, technology & transformation for an unknown future**. Proceedings ASCILITE Sydney, p. 75-86, 2010.

BRASIL. Curso de Formação de Monitores do Telecentros.Br **Manual Operacional da Rede Nacional de Formação para Inclusão Digital**. Ministério do Planejamento. Secretaria de

BRASIL. **Decreto n.° 6991**, de 27 de outubro de 2009, Institui o Programa Nacional de Apoio à Inclusão Digital nas Comunidades Telecentros.BR, no âmbito da política de inclusão digital do Governo Federal, e dá outras providências. Diário Oficial [da] Republica Federativa do Brasil, Brasília, DF, n. 206, Seção 1, pág. 3, 2009.

BRASIL. **Manual Operacional da Rede Nacional de Formação para Inclusão Digital**. Ministério do Planejamento. Secretaria de Logística e Tecnologia da Informação (SLTI) - Assessoria de Inclusão Digital. Brasília: SLTI, 2011. Disponível: http://www.slideshare.net/telecentrosbr/documento-orientador-da-rededе-formao-janeiro-2011. Acesso em 10 mar. 2013. 2011.

BREUER, C., HALLMANN, K., WICKER, P., & FEILER, S. **Socio-economic patterns of sport demand and ageing**. European Review of Aging and Physical Activity, 7(2), 61-70, 2010

CRAIN, S. P.; ZHOU, K.; YANG, S. H.; ZHA, H. Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. **In: Mining text data**. Springer US. pp. 129-161, 2012

DARELLI, L**. Telecentro como instrumento de inclusão digital para o e-gov brasileiro**. Master's thesis. Programa de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina, 2002.

DE BRITO, S. R., DA SILVA, A. D. S., MARTINS, D. L., DA ROCHA, C. A. J., COSTA, J. C. W. A., & Francês, C. R. L. **Brazilian Government's Training Network for Digital Inclusion: Analysis of Strategies for Improving Interactivity**. Handbook of Research on Enterprise 2.0: Technological, Social and Organization Dimensions. IGI Global, 2013a.

DE BRITO, S. R., DA SILVA, A. D. S., MARTINS, D. L., VIJAYKUMAR, N. L., DA ROCHA, C. A. J., Costa, J. C. W. A., & Francês, C. R. L. **Employing online social networks to monitor and evaluate training of digital inclusion agents. Social Network Analysis and Mining**, 3(3), 497-519. 2013b.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. **From Data Mining to Knowledge Discovery**: An Overview, in Advances in Knowledge Discovery and Data Mining, R. Uthurusamy, eds., MIT Press, Cambridge, Mass., pp. 1-36. 1996.

FELDMAN, R.; SANGER J. **The Text Mining Handbook** – Advanced Approaches in Analyzing Unstructed Data. Cambridge University Press, New York, 2007.

FRASCARELI, A.M.F. AND PIMENTEL, E.P. **Aplicando técnicas de bibliometria, mineração de texto e vizualização na identificação de temas e tendências de pesquisa em e-learning**, anais do XXIII SBIE, p. 26-30, 2012.

GOSAIN, Anjana; KUMAR, Amit. Analysis of health care data using different data mining techniques. **In: Intelligent Agent & Multi-Agent Systems**, 2009. IAMA 2009. International Conference on. IEEE, 2009. p. 1-6.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. **Avaliação de desempenho de estudantes em cursos de educação a distância utilizando Mineração de Dados**. Anais do XXXII Congresso da Sociedade Brasileira de Computação. 2012a.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. **Estimativa de Desempenho Acadêmico de Estudantes**: Análise da Aplicação de Técnicas de Mineração de Dados em Cursos a Distância. Revista Brasileira de Informática na Educação, Vol. 22, N. 1, pp. 45-55, 2014.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. **Previsão de desempenho de estudantes em cursos EAD utilizando Mineração de Dados**: uma estratégia baseada em séries temporais, Anais do XXIII Simpósio Brasileiro de Informática na Educação – SBIE, Rio de Janeiro, 2012b.

HÄMÄLÄINEN, W.; VINNI, M. Classifiers for Educational Data Mining. In: Romero et al. **Handbook of Educational Data Mining**. Flórida, CRC Press, p. 57-71, 2011.

HAN, J., & KAMBER, M. **Data Mining**: Concepts and Techniques. University of Illinois at Urbana-Champaign, 2006

HEARST, M. A. **Untangling text data mining**. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, p. 3-10, 1999.

Hu, H. J., Harrison, R. W., Tai, P. C., Pan, Y. Understandable learning machine system design for Transmembrane or Embedded Membrane segments prediction. **International journal of data mining and bioinformatics**, v. 5, n. 1, p. 38-51, 2011.

HU, X.; LIU, H. Text analytics in social media. **In: Mining text data.** Springer US, 2012. p. 385-414.

JOLLIFFE, I. Principal component analysis. John Wiley & Sons, Ltd, 2002.

KAMPFF, A.J.C. FERREIRA, V. H.; REATEGUI, E.; de LIMA, J. V. Identificação de perfis de evasão de Evasão e Mau Desempenho para Geração de Alertas num Contexto de Educação

a Distância**. Revista Latino-Americana de Tecnologia Educativa**, Vol. 13(2), pp. 61- 76, 2014.

KAMPFF, A.J.C. **Mineração de Dados Educacionais para a Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente.** Tese de Doutorado. Programa de Pós-Graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009

KOHONEN, T. **Self-Organizing Maps**. Springer, Berlin, Heidelberg. 1965.

KOVALERCHUK, B., VITYAEV, E.: **Data Mining for Financial Applications**. In: O. Maimon, L. Rokach (Eds.): The Data Mining and Knowledge Discovery Handbook. Springer, pp. 1203-1224, 2010. Second edition.

LAW, E. L.-C. et al. **Understanding, scoping and defining user experience**: a survey approach. In: CHI - International Conference On Human Factors In Computing Systems, 27. Anais eletrônicos. Nova Iorque: ACM, 2009. p.719-728.

LOGÍSTICA E TECNOLOGIA DA INFORMAÇÃO (SLTI) - Assessoria de Inclusão Digital. Brasília: SLTI, 2010.

LOPES, M.C.S. **Mineração de dados textuais utilizando técnicas de clustering para o idioma português**. Master's thesis. Universidade Federal do Rio de Janeiro, 2004.

MACFADYEN, L.P.; DAWSON, S. **Mining LMS Data to Develop an 'Early Warning System' for Educators**: A Proof of Concept. Computers & Education, no. 54, p. 588-599, 2010.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Evaluation of unranked retrieval sets. Introduction to Information Retrieval** , v. 7, 2009.

MARTINS, D.; FLAUZINO, R.; DIAS, S. **Gestão em rede e design Instrucional: um relato de experiência do Programa Telecentros.BR**. 17º Congresso Internacional de Educação a Distância, Manaus, 2011.

MCAULEY, A.; STEWART, B.; SIEMENS, G.; CORMIER, D. **The MOOC model for digital practice**.. 2010. Disponível em < http://www.davecormier.com/edblog/wp-content/uploads/MOOC_Final.pdf >. Acesso em 11. Ago. 2015

MOODLE. Site, 2005. Disponível em http://moodle.org . Acesso em 27 abr 2015.

MOORE M. G. **Three Types of Interaction**. The American Journal of Distance. 1989.

MOSTOW, J., Et Al. An **educational data mining tool to browse tutor–student interactions**: Time will tell! In: Proceedings of the workshop on educational data mining, pp. 15–22, 2005.

OKONKWO, R. O.; ENEM, F. O. Combating crime and terrorism using data mining techniques. In: **10th International conference IT people centred development**, Nigeria Computer Society, Nigeria. 2011.

PINHEIRO, M. F., NETO, L. C. F., DE SÁ JUNIOR, H. N., DA MATA, E. C., JACOB JR, A. F., & DE LIMA SANTANA, Á. (2014a). **Identificação de Grupos de Alunos em Ambiente Virtual de Aprendizagem Utilizando Análise de Log Baseada em Clusterização**. In Anais do XLII Congresso Brasileiro de Educação em Engenharia.

PINHEIRO, M. F., NETO, L. C. F., DE SÁ JUNIOR, H. N., DA MATA, E. C., JACOB JR, A. F., & DE LIMA SANTANA, Á. (2014b). **Identificação de Grupos de Alunos em Ambiente Virtual de Aprendizagem: Uma Estratégia de Análise de Log Baseada em Clusterização**. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação (Vol. 3, No. 1).

PNAD. **Pesquisa Nacional por Amostra de Domicilio**. Instituto Brasileiro de Geografia e Estatística. Disponível em <http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2013/>. Acesso em 19 Jul 2015.

QUINLAN, J.R. **C4.5 Programs for Machine Learning**, San Mateo, CA: Morgan Kaufmann, 1992.

QUINLAN, J.R. **Discovering rules by induction from large collections of examples**. In D. Michie (Ed.), Expert systems in the microelectronic age. Edinburg University Press, 1979.

RABBANY, R.K.; TAKAFFOLI, M.; ZAÏANE, O.R. **Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques**. In Proceedings of the Fourth International Conference on Educational Data Mining, p. 22-30, 2011.

RAMOS, J. Using **TF-IDF to determine word relevance in document queries**. In: Proceedings of the first instructional conference on machine learning. 2003.

REDE DE FORMAÇÃO TELECENTROS.BR, **Indicadores Web Telecentros.Br – Redes Sociais**. Disponível em < http://pt.slideshare.net/telecentrosbr/telecentrosbr-indicadores-web-resumo?qid=0f631951-58ee-4e5d-bb59-990591007c56&v=default&b=&from_search=1>. Acesso em 11 Ago. 2015. 2011b.

REDE DE FORMAÇÃO TELECENTROS.BR, **Monitoramento da Formação Telecentros.BR**. Disponível em < http://pt.slideshare.net/telecentrosbr/relatrio-monitoramento-da-formao-abril-2011-telecentros-br>. Acesso em 11 Ago. 2015. 2011a.

REZENDE, S. O. **Sistemas inteligentes**: fundamentos e aplicações. Barueri, Manole, 2005.

RICARTE, I. L. M.; JUNIOR, G.R. F.. **A Methodology for Mining Data from Computer-Supported Learning Environments**. Informática na educação: eoria & prática, v. 14, n. 2, 2011.

RIEDO, C., PEREIRA, E., WASSEM, J., GARCIA, M. **O desenvolvimento de um mooc (massive open online course) de educação geral voltado para a formação continuada de professores: uma breve análise de aspectos tecnológicos, econômicos, sociais e pedagógicos.** SIED: EnPED-Simpósio Internacional de Educação a Distância e Encontro de Pesquisadores em Educação a Distância. 2014.

RODRIGUES, R. L.; RAMOS, J.L.C; SILVA, J.C.S; GOMES, A. S. **A literatura brasileira sobre a mineração de dados educacionais**. Anais dos Workshops do Congresso Brasileiro de Informática na Educação. Vol. 3. No. 1. 2014.

ROMERO, C.; VENTURA, S. Data **Mining in course management systems**: Moodle case study and tutorial. In: Computers & Education, Elsevier, pp. 368-384, 2008.

ROMERO, C.; VENTURA, S. **Educational data mining**: A review of the state of the art. In: IEEE Transactions on Systems, Man, and Cybernetics – Applications and Reviews, vol. 40, no. 6, pp. 601-618, 2010.

ROMERO, C.; VENTURA, S. **Educational data mining**: A survey from 1995 to 2005. In: Expert Systems with Applications. Vol. 33, Elsevier, pp. 135-136, 2007.

ROMERO, Cristobal; VENTURA, Sebastian. **Data mining in education**. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 3, n. 1, p. 12-27, 2013.

SACHIN, R. B. ARAHATE;; VIJAY, M. S.HELAKE. **A survey and future vision of data mining in educational field**. In: Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on. IEEE, 2012. p. 96-100.

SALTON, G.; MCGILL, M. **Introduction to Modern Information Retrieval**. McGraw Hill, New York, 1983.

SALTON, G.; WONG, A.; YANG, C. **A vector space model for automatic indexing.** Communications of the ACM, v. 18, n. 11, p. 613-620, 1975. (Third Extended Edition 2001).

SANJEEV, P.; ZYTKOW, J. M. **Discovering enrollment knowledge in university databases.** In: KDD, pp. 246-251, 1995.

SANTANA, L. C.; MACIEL, A. M. A.; RODRIGUES, R. L. **Avaliação do perfil de uso no ambiente Moodle utilizando técnicas de Mineração de Dado**s. Anais do III Congresso Brasileiro de Informática na Educação – CBIE, 2014.

SCHROECK, M.; SHOCKLEY, R.; SMART, J.; ROMERO-MORALES; D.; TUFANO, **Analytics**: The real-world use of big data. IBM Global Business Services, Somers, 2012.

SHABALIN, A. **Visuals and Animations**. 2007. Disponível em < http://shabal.in/visuals/kmeans/2.html >. Acesso 19 Jul 2015.

SILVA, A.; DE BRITO, S.; VIJAYKUMAR, N. L.; MARTINS, D.; DA ROCHA, A.; COSTA, J. C. W. A.; FRANCÊS, C. R. L. **Análise de Redes Sociais para avaliação e monitoramento de programas de treinamento em larga escala baseados no uso de ambientes de**

aprendizagem e redes sociais online. II Brazilian Workshop on Social Network Analysis and Mining, 2013.

SILVA, L.; M**ORINO, A.; SATO, T**. Prática de Mineração de Dados no Exame Nacional do Ensino Médio. Anais do 3º Congresso Brasileiro de Informática na Educação – CBIE, 2014.

VALENTIM, G.; GUZZI, D.; MARTINS, D.; NOGUCHI, N.; SOARES, I.; MANFREDI, M. **Relato do processo de monitoramento da formação para inclusão digital do programa Telecentros.Br**. 17º Congresso Internacional de Educação a Distância, Manaus, 2011.

VITYAEV, Evgenii; KOVALERCHUK, Boris. **Relational methodology for data mining and knowledge discovery**. In: Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on. IEEE, 2005. p. 725-729.

WANG, J.T., ZAKI, M. J., TOIVONEN, H. T., & SHASHA, D. **Introduction to data mining in bioinformatics**. Springer London, pp. 3-8, 2005.

WITTEN, H.; FRANK, E.; HALL, M. A. **Data Mining**: Practical Machine Learning Tools and Tech-niques. San Francisco: Morgan Kaufmann, 3 ed, 2011.

WU LF *et al.* **Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters.** Nat. Genet; pp.255-265, 2002.

ZAÏANE, O.; XIN, M., HAN, J. **Discovering Web access patterns and trends by apllying OLAP and data mining technology on web logs**. In: Advances in Digital Libraries, pp. 19-29, 1998.

ANEXO A – ARTIGO ACEITO EM CONGRESSO

IDENTIFICAÇÃO DE GRUPOS DE ALUNOS EM AMBIENTE VIRTUAL DE APRENDIZAGEM: UMA ESTRATÉGIA DE ANÁLISE DE LOG BASEADA EM CLUSTERIZAÇÃO

Márcia F. Pinheiro[1], Luiz Cortinhas F. Neto1[1], Haroldo Nazaré de Sá Junior[1] Eulália C. da Mata[1], Antonio F. L. Jacob Jr. [1], Ádamo de Lima Santana[1]

[1]Instituto de Tecnologia – Universidade Federal do Pará (UFPa)

Caixa Postal 479 – 66075-110 – Belém –PA – Brasil

{eng.marciafontes, luizcf14, hnsj86, eucmata}@gmail.com,

{jacobjr, adamo}@ufpa.br

Abstract. *Have information about the learning process is of utmost importance for educators and students, as it allows to support decision making and reflection regarding teaching methodologies and contents used and performance of students. In this study were used Mining techniques in educational data in order to find groups of students from a large-scale educational program by analyzing a Distance Learning platform logs. The results obtained allow making qualitative and quantitative analysis of the use of the Moodle platform to groups of students found.*

Resumo. *Ter informações a respeito do processo de aprendizagem é de extrema importância para educadores e alunos, pois permite apoiar a tomada de decisão e reflexão quanto às metodologias aplicadas no ensino, bem como conteúdos utilizados e desempenho de alunos. Neste trabalho, foram utilizadas técnicas de mineração de dados educacionais com o objetivo de encontrar grupos de alunos de um programa educacional de larga escala através da análise de log de uma plataforma de Educação a Distância como parte de pesquisa motivada pela verificação de aproveitamento dos alunos na formação. Os resultados obtidos permitem fazer análise qualitativa e quantitativa do uso da plataforma pelos dos grupos de alunos encontrados.*

1. INTRODUÇÃO

Com o advento das Tecnologias da Informação e Comunicação (TICs), Internet, dos sistemas de informação para Web e dos recursos multimídia, surgiram diversas ferramentas e plataformas com a proposta de Web 2.0.

No campo da educação, a utilização desses sistemas de ensino eletrônico – e-learning – proporcionam novas possibilidades de aprendizagem na metodologia de ensino. Esses sistemas são denominados como Ambientes Virtuais de Aprendizagem (AVAs) por serem plataformas

que centralizam ferramentas, mas que permitem interação entre os usuários de forma assíncrona e síncrona através de Wiki, chat, fórum, etc.

Os AVAs armazenam todas as interações dos usuários dentro da plataforma, informações do desempenho do aluno, podendo guardar informações como quais atividades um estudante participou, materiais que foram lidos e escritos, testes ao qual foi submetido, chats que o mesmo participou, páginas acessadas na plataforma, etc. [Mostow et al., 2005], assim como informações pessoais sobre usuário, tais como perfil, em sua base de dados.

Dentre os AVAs, temos o Moodle, que é um sistema *open source* desenvolvido pelo australiano Martin Dougiamas para uso acadêmico. Engloba diversas ferramentas web 2.0 para facilitarem a interação entre os usuários [Mata et al., 2010] e é a plataforma mais utilizada no mundo em educação a distância [CAPTERRA, 2014].

De maneira geral, os AVAs armazenam uma grande quantidade de informação o qual é muito valiosa para analisar o comportamento dos estudantes [Mostow and Beck, 2006] e que pode ser obtida através de Mineração de Dados, também conhecida como Descoberta do Conhecimento em Base de Dados (DCBD), da expressão em inglês Knowledge Discovery in Databases (KDD), que é a extração automática de padrões implícitos e interessantes a previamente desconhecidas e potencialmente úteis a partir de grandes volumes de dados [Klösgen and Zytkow, 2002][WITTEN, FRANK and HALL, 2005]. Para encontrar padrões nesses dados são envolvidos a Mineração de Dados (*Data Mining*), extração do conhecimento, descoberta da informação e padrão de processamento dos dados [Fayyad, 1996]. Baker e de Carvalho (2010) complementam que a Mineração de Dados permite descobrir novas informações no processo de identificação de relações entre dados que podem produzir novos conhecimentos e gerar novar descobertas científicas. Informações estas que podem ser muito úteis para tomada de decisão.

KDD tem sido aplicada em diversas áreas do conhecimento, como por exemplo, em finanças [KOVALERCHUK and VITYAEV, 2005], bioinformática [HU, 2011], combate ao crime e terrorismo [OKONKWO AND ENEM, 2011], saúde [GOSAIN and KUMAR, 2009], esportes [WICKER AND BREUER, 2010], etc. A Informática na Educação é uma linha de pesquisa que tem sido consolidada, como apresenta o trabalho de Frascareli e Pimentel (2012),

e tem sido um tema estudado por diversos pesquisadores da área, em particular da Inteligência Artificial Aplicada à Educação [BARKER, ISOTANI and De Carvalho, 2011].

KDD tem sido utilizada com o intuito de investigar perguntas cientificas na área de educação como, por exemplo, quais são os fatores que afetam a aprendizagem? Como desenvolver sistemas educacionais mais eficazes? Ou ainda a relação da abordagem pedagógica e o aprendizado do aluno, estas informações podem ser úteis não somente para os educadores, mas também aos próprios alunos, uma vez que pode ser orientada para diferentes fins por diferentes participantes no processo. Dentro deste contexto surgiu a Mineração de Dados Educacionais (do inglês, Educational Data Mining - EDM), que 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, desta maneira é possível compreender de maneira mais eficaz e adequada os alunos, como estes aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem.

Este trabalho propõem como metodologia à aprendizagem de log (JANSEN, 2006) adaptada para plataforma Moodle utilizada em grande escala para cursos de incentivos públicos e privados semi-presenciais brasileiros, aplicando clusterização através da abordagem de log para encontrar grupos de alunos na plataforma bem como o caminho médio de cada grupo na plataforma, como trabalho inicial de verificação de aproveitamento dos alunos na formação.

Na Seção 2 são apresentados trabalhos que utilizaram técnicas de EDM; a Seção 3 apresenta a contextualização do domínio o qual esse trabalho realizou a pesquisa; Na Seção 4 são apresentados os experimentos realizados; Na Seção 5 são apresentados os resultados obtidos e a discussão dos mesmos e na Seção 6 são feitas as conclusões e possíveis trabalhos futuros.

2. TRABALHOS RELACIONADOS

Como área de pesquisa em crescimento [Frascareli and Pimentel, 2012], diversas pesquisas têm sido feitas no campo da EDM utilizando dados de AVAs. Por exemplo, Gottardo, Kaestner e Noronha (2012) aplicaram EDM utilizando classificação em busca da obtenção de modelos para inferir e prever o desempenho dos estudantes em um curso de Educação a Distância (EAD) a partir de dados coletados no AVA Moodle.

Dentre os diversos métodos utilizados em EDM, muitos deles originalmente são da área de mineração de dados, como predição, mineração de relações e clusterização, por exemplo. A técnica de agrupamento ou clusterização tem como objetivo principal achar dados que se agrupam naturalmente, classificando os dados em diferentes grupos e/ou categorias. Estes grupos e categorias não são conhecidos a priori. Através de técnicas de agrupamento são automaticamente identificados através da manipulação das características dos dados.

Vários trabalhos envolvendo técnicas de EDM foram desenvolvidos com o intuito de avaliar o comportamento dos usuários de AVAs e propor melhorias na organização de AVAs ou conteúdos dos respectivos cursos. Ricarte e Falci Júnior (2011) aplicaram clusterização utilizando os algoritmos K-means e Mapas Auto-organizáveis (do inglês, Self-Organized Map - SOA) em uma base de *logs* coletada a partir do uso de um AVA na UNICAMP para encontrar grupos de estudantes com comportamento semelhante com o intuito de oferecer retorno a autores e tutores sobre o uso dos conteúdos disponibilizados, bem como oferecer a estudantes sobre seu próprio uso dos recursos do ambiente.

Swedan (2012) aplicou clusterização utilizando K-means para encontrar os diferentes grupos de alunos quanto ao nível de participação no AVA Moodle a partir do uso dos recursos deste. O autor propôs ainda uma ferramenta estatística para os educadores monitorarem os alunos e seus respectivos comportamentos de aprendizagem a partir dos *logs* coletados do Moodle. Analogamente, Bovo et al (2013) utilizaram clusterização para encontrar diferentes grupos de alunos e se estes grupos apresentam diferenças qualitativas e quantitativas.

Romero, Ventura e García (2008) verificaram que as ferramentas de mineração de dados disponíveis são demasiadamente complexas para educadores, que precisam de interfaces mais fácies de analisar os resultados da mineração para avaliar o processo de aprendizagem. Desta forma é mais provável o administrador do AVA aplicar técnicas de mineração, a fim de produzir relatórios para os educadores que, em seguida, usarão esses relatórios para tomar decisões sobre como melhorar a aprendizagem do aluno e dos cursos *online.*

Neste trabalho a técnica de EDM utilizada é a Clusterização aplicando o algoritmo *K-means* a uma base de *logs* reais coletada do Moodle durante um curso EAD, com o objetivo de encontrar grupos de usuários com comportamento similar para avaliar qualitativamente e quantitativamente estes grupos.

3. CONTEXTUALIZAÇÃO DO PROBLEMA E DESCRIÇÃO DA BASE DE DADOS

No Brasil, dentre as políticas públicas para favorecimento da inclusão digital, têm-se investido na criação e na utilização de centros tecnológicos comunitários, ou telecentros, onde o acesso público às TICs é disponibilizado para as comunidades menos privilegiadas a um custo mínimo ou isento de custos. Neste contexto, além da implantação dos telecentros para acesso à Internet, têm-se a formação de agentes para inclusão digital como aspecto crítico. A proposta envolveu um programa de formação em larga escala que requereu mecanismos de controle e monitoramento para gestores e beneficiadores desse programa, a Rede Nacional de Formação para Inclusão Digital – Rede Telecentros BR [Silva, 2013].

A Rede Telecentros BR até o ano de 2011 contava com a participação de cinco polos regionais (um polo para cada região do País), dois polos estaduais (nos estados de São Paulo e Ceará) e um polo nacional. Sob a responsabilidade dos polos regionais estava a formação dos agentes de inclusão digital (que são os monitores dos telecentros), gestores de telecentros (responsáveis pela administração do telecentro), tutores (que atuavam na formação dos monitores) e supervisores de tutoria (responsáveis pela supervisão e acompanhamento do trabalho dos tutores) [Silva, 2013].

No período de fev/2010 a dez/2012, os membros dos polos de formação se articularam para construir e aplicar o Curso de Formação de Monitores dos Telecentros e a ativação das redes sociais de agentes de inclusão social atuantes nas comunidades. O projeto de formação dos agentes de inclusão digital (monitores) contemplou a oferta de um curso de 480 horas, disponibilizado na plataforma Moodle (Rede Telecentros.BR, 2013) e dividido em dois módulos: 80 horas para um uma breve apresentação dos conteúdos da formação e 400 horas com foco específico no desenvolvimento de projetos comunitários e aprofundamento dos conteúdos. Para o desenvolvimento dos projetos comunitários, os agentes de inclusão digital percorreram de acordo com seus interesses e necessidade, sem percurso pré-definido, diferentes temas: comunicação comunitária, redes, cultura digital, comunidade, telecentros. No desenvolvimento do curso, os agentes de inclusão digital contaram com o apoio de tutores e supervisores de tutores dos diversos polos regionais [Silva, 2013].

3.1. Plataforma Moodle

O Moodle é o Ambiente Virtual de Aprendizagem (AVA) selecionado para o desenvolvimento do Curso de Formação de Monitores do Telecentros.BR. A Rede Nacional de

Formação para Inclusão Digital produziu o curso de formação de monitores do Telecentros.BR com objetivo de propiciar o desenvolvimento de habilidades no uso de tecnologias da informação e comunicação para dar condições de transformações sociais na comunidade. O curso foi estruturado em dois eixos pedagógicos: acessos a conteúdos e atividades formativas; elaboração e implementação de projetos comunitários. Ofertado em duas fases: a primeira englobava conhecer o ambiente virtual e abordagem prévia do conteúdo que será aprofundado na fase dois que aborda principalmente o projeto comunitário, onde cada tema estudado poderá servir de apoio para realização de ações com a comunidade. Entre os temas abordados: telecentros, comunidade, comunicação comunitária, inclusão digital, redes e cultura digital.

No ambiente virtual de aprendizagem – Moodle foi organizado o curso para formação dos monitores em: Fase 1: Ambientação e Voo Rasante; Fase 2: Projeto Comunitário. Também foi necessário realizar um curso para formação de tutores que atuaram auxiliando e orientando os monitores durante a formação do telecentros.br. No curso dos tutores foi abordado educação a distância, articulação social e inclusão digital. Os tutores foram selecionados com base no conhecimento sobre tecnologia, valorizando a experiência em inclusão digital.

O processo de formação foi realizado em grupo de “n” monitores para um tutor, que fizeram o acompanhamento para o primeiro acesso a plataforma, incentivaram a participação no ambiente virtual e no desenvolvimento do projeto comunitário, assim como realizaram a avaliação dos monitores mensalmente. Os tutores tiveram o acompanhamento e auxilio dos supervisores de tutoria no processo de formação. Os supervisores eram pessoas que fizeram parte dos polos regionais, facilitadores da formação, acompanhavam os tutores e monitores no processo de formação, realizaram também avaliação das turmas com base nas informações levantadas na plataforma moodle e com as informações repassadas pelos tutores. A coordenação pedagógica realizou acompanhamento dos supervisores e tutores com base em informações obtidas na plataforma e no sistema de avaliação.

4. EXPERIMENTOS REALIZADOS

Para a realização deste trabalho foi utilizada a base de dados do Moodle utilizado na Rede Telecentros BR por aproximadamente mil e quatrocentos alunos (monitores) no curso de

formação para inclusão digital promovido pela Rede Nacional de Formação para Inclusão Digital, a qual foi responsável pela construção, manutenção e aplicação do referido curso.

A coleta de dados ocorreu de maneira implícita através do uso e interação dos monitores na plataforma Moodle e armazenamento na base de dados da plataforma. O Moodle armazena todas as interações dos usuários na plataforma em forma de logs: cada clique que o usuário realiza na plataforma para fins de navegação, bem como detalhes das atividades que os estudantes participaram. Uma vantagem de registrar as atividades de um usuário em uma plataforma em forma de log é que grande grandes volumes de dados podem ser armazenados automaticamente [Rogers, Sharp and Preece, 2011].

O Moodle armazena os logs em uma base de dados relacional, com cerca de 145 tabelas inter-relacionadas, porém neste trabalho as informações utilizadas encontravam-se somente na tabela *mdl_log* responsável pelo armazenamento de eventos do sistema. O pré-processamento consistiu na limpeza das informações das demais tabelas do Moodle em conjunto com o filtro para eliminar alunos excluídos e administradores do sistema procurando mitigar a interferência destes grupos de usuários.

Neste trabalho foi utilizado o algoritmo *K-means* que é um método de clusterização (organização dos objetos similares, em algum aspecto, em um grupo) pertencente à classe dos algoritmos de aprendizados de máquina. Tem como finalidade dividir um determinado número de objetos em áreas chamadas de *cluster*. Estas áreas são constituídas por uma coleção de objetos que são similares entre si e diferentes dos objetos pertencentes aos outros *clusters*. A distância de cada objeto para cada *cluster* vai determinar em que *cluster* o objeto ficará alocado. Cada objeto pertence ao *cluster* no qual possui o elemento central (centróide) mais próximo deste objeto

A partir da etapa de pré-processamento a clusterização é permitida usando o algoritmo K-means para duas dimensões onde são tratadas as colunas: “course” e “userid”, contando o número de acessos de cada usuário para cada curso, também são filtrados somente eventos dentro da categoria “course”. Após a composição do *dataset* para o algoritmo de clusterização o mesmo será executado para: k=5, pois são estabelecidos pelo programa quatro níveis de avaliações qualitativas: Excelente, Bom, Regular e Insuficiente, contudo foi acrescentada uma

classe para que *outliers* (valores fora do esperado) possam ser minimizados nas outras classes (FAYYAD, U et al, 1996).

Os resultados do *K-means* são grupos definidos que passam pela análise de quantitativos de alunos, acessos e cursos.

A partir do mesmo é possível inferir características comuns entre grande número de usuários no mesmo grupo e ate mesmo o grupo dos usuários mais experientes na plataforma, chamados monitores.

5. Resultados

Para o pré-processamento o numero de linhas da base foram resumidas para 13283 linhas, pois as linhas de log ficaram restritas as características citadas na metodologia.

Para o *K-means* foram encontrados os 5 clusters (denominados K1, K2, K3, K4, K5) com as seguintes características como mostra a Tabela 1:

Tabela 1. Clusters encontrados

	K1	K2	K3	K4	K5
Usuários	5	32	183	825	3251
Acessos	228728	598867	1197017	2024758	1152300
Cursos	50	50	50	49	50
Média de Acesso por Usuários	45745,6	18714,59375	6541,076502732	2454,252121212	354,44478622
Média de Acesso por Curso	914,912	374,291875	130,821530055	50,086777984	7,088895724
Ação mais utilizada	View	View	View	View	View

	93930	266186	614599	1183165	93930
Recursos mais visualizados	Error Login 59430	Course 153197	Forum 294337	Course 513217	Course 310376
Caminhos Médios	145	163	167	137	133

É importante ressaltar o tempo de execução em 5 minutos e 55 segundos, para o pré-processamento e k-means. Todos os procedimentos foram executados no mesmo computador com os seguintes requisitos de hardware: Processador Core i7-3632QM @ 2.20Ghz, Memoria de 8GB @1333Mhz e sistema operacional Linux distribuição Ubuntu 14.04.

5.1. Discussão dos Resultados

Os resultados demonstraram grupos distintos de usuários a partir da clusterização do *log* da plataforma Moodle. É possível reconhecer com base nos resultados a distância entre o maior e menor cluster isso ocorre, pois no caso do cluster K1, pelo pequeno número de usuários e grande número de acessos com erro no Login o grupo demonstra ser formado provavelmente por usuários com comportamento fora do esperado (*outliers*), No cluster K2, identificamos pelo número de usuários e a grande quantidade de acessos os usuários responsáveis pela capacitação dos monitores, chamados Tutores. Nos clusters K3, K4 e K5 encontramos números de usuários distintos, porém com a média de acesso geral semelhante e com a média por usuário decrescente, respectivamente, então, é possível inferir que estão localizados alunos, chamados monitores, com bom acesso ao sistema no cluster K3, monitores com aceso regular ao sistema no cluster K4 e monitores com um baixo número de acesso na plataforma no cluster K5.

A partir da análise dos recursos mais utilizados, podemos inferir que os monitores do cluster K3 realmente foram mais participativos pelo uso do recurso *Forum*, enquanto que os monitores do cluster K4 tiverem uma maior participação nos materiais dos seus respectivos cursos, assim como os usuários do cluster K5 tiveram participação razoável em seus respectivos cursos. Entende-se por caminho médio quais as ações que foram mais utilizadas junto a recursos na plataforma por cada *cluster* encontrado, estes caminhos médios

foram levantados em busca de encontrar as ações em comum dos usuários de cada cluster na plataforma.

Para os clusters que representam os monitores, que são alunos na plataforma, representados por K3, K4 e K5, obteve-se o uso dos recursos disponíveis para os referidos clusters, respectivamente, como mostram a Figuras 1, 2 e 3.

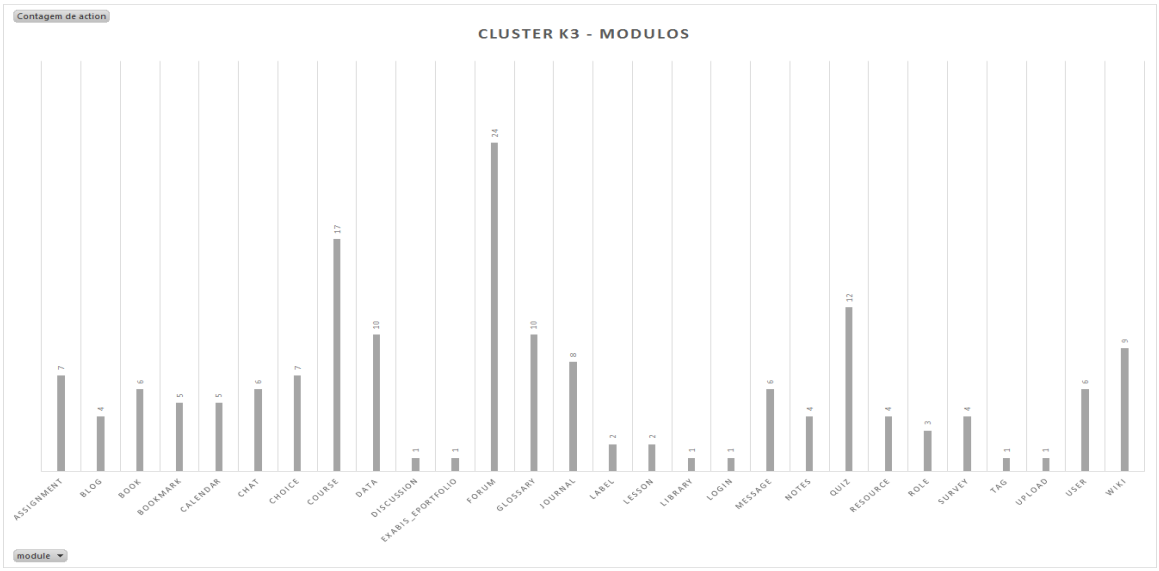

Figura 1. Uso dos recursos pelo cluster K3

Figura 2. Uso dos recursos pelo cluster K4

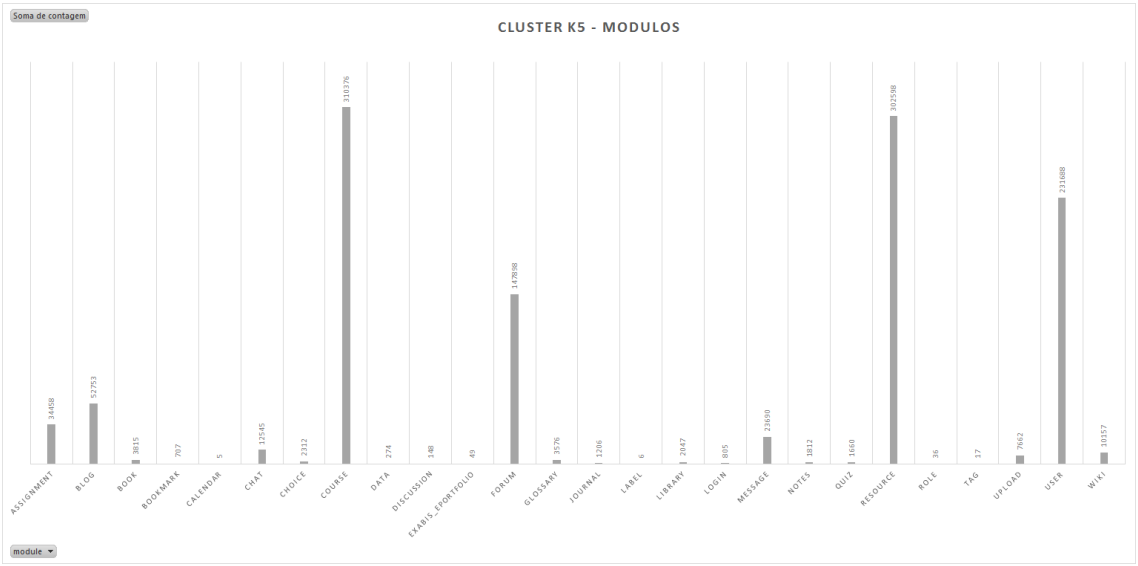

Figura 3. Uso dos recursos pelo cluster K5

Percebe-se que o recurso que mais utilizado na plataforma pelos alunos com acesso regular e com baixo acesso, clusters K4 e K5, durante a formação foi o "Curso", seguido pelo "Recursos" (Arquivos, Mídias, etc.). Enquanto que o recurso de "Fórum" foi o mais utilizado pelos alunos com maior número de acessos à plataforma, seguido pelo "Curso". Ferramentas importantes de comunicação como Chat, por exemplo, foram muito pouco utilizadas por estes alunos.

5. Conclusões e trabalhos futuros

Este trabalho apresentou como proposta exploratória (Gil, 2008) a utilização de técnicas de EDM com o intuito de encontrar grupos de alunos de um programa educacional de larga escala através da análise de log de uma plataforma de Educação a Distância. Com base nas técnicas que são reportadas pela literatura, aplicou-se o algoritmo *K-means* em busca do agrupamento de alunos com comportamentos semelhantes na plataforma utilizando a análise do caminho médio para encontrar as ações em comum dos usuários de cada *cluster* dentro da plataforma Moodle. Os resultados encontrados são de extrema importância para o gerenciamento dos cursos no Moodle e para as análises quantitativa e qualitativa feitas no programa Telecentros BR.

Como trabalho futuro pretende-se correlacionar os grupos de monitores (alunos) encontrados neste trabalho com seus respectivos caminhos médios com suas respectivas avaliações realizadas em outro sistema pelos tutores (educadores). Para tal será necessário a

aplicação técnicas de KDD utilizando algoritmos de Mineração de Texto, para a obtenção das notas dos monitores, pois as avaliações são qualitativas e não quantitativas. Conseguinte pretende-se utilizar outras técnicas de Inteligência Artificial para classificação dos alunos segundo suas notas.

REFERÊNCIAS

Baker, R.S.J.D. (2010) Data Mining for Education. In: McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education. Oxford, UK: Elsevier, 3ed.

Baker, R.S.J.D., Isotani, S. and, de Carvalho, A.M.J.B. (2011) "Mineração de Dados Educacionais: Oportunidades para o Brasil". Revista Brasileira de Informática na Educação, vol. 19, no. 2, p. 2-13.

Bovo, A., Sanchez, S., Héguy, O. and Duthen, Y. (2013) "Clstering Moodle Data as a Tool for Profiling Students", In: Second International Conference on e-Learning and e-Technologies in Education (ICEEE), p. 121-126

Capterra (2014). The Top 20 Most Popular Lms Soft Ware Solutions. Http://Www.Capterra.Com/Top-20-Lms-Software-Solutions.Uzjrp7wncsp.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) From Data Mining to Knowledge Discovery: An Overview, in Advances in Knowledge Discovery and Data Mining, R. Uthurusamy, eds., MIT Press, Cambridge, Mass., pp. 1-36.

Frascareli, A.M.F. and Pimentel, E.P (2012) "Aplicando Técnicas de Bibliometria, Mineração de Texto e Vizualização na Identificação de Temas e Tendências de Pesquisa em *e-Learning*", Anais do XXIII SBIE, p. 26-30.

MACQUEEN, J. On convergence of k-means and partitions with minimum average variance. In: Annals of Mathematical Statistics. IMS BUSINESS OFFICE-SUITE 7, 3401 INVESTMENT BLVD, HAYWARD, CA 94545: INST MATHEMATICAL STATISTICS, 1965. p. 1084-&.

Gil, A. C. (2008) "Métodos e Técnicas de Pesquisa Social", São Paulo, Atlas.

Gosain, A and Kumar, A, (2009) "Analysis of health care data using different data mining techniques", International Conference on Intelligent Agent & Multi-Agent Systems (IAMA), vol. 1, no. 6, pp. 22-24

Gottardo, E. Kaestner, C. and Noronha (2012) “Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: Uma Estratégia Baseada em Séries Temporais”, Anais do XXIII SBIE.

Hu, X. (2011) “Data mining and its applications in bioinformatics: Techniques and methods”, IEEE International Conference on Granular Computing (GrC), vol.3 (3): , no. 3, p. , 8-10.

Klösgen, W. and Zytkow, J. (2002) Handbook of data mining and knowledge discovery. New York: Oxford University Press.

Kovalerchuk, B. and Vityaev, E. (2005) “Data Mining for Financial Applications”, Data Mining and Knowledge Discovery Handbook. Springer US, p. 1203-1224.

Mostow, J. and Beck, J. (2006) Some useful tactics to modify, map and mine data from intelligent tutors. Journal Natural Language Engineering, vol. 12, pp. 195-208.

Mostow, J., Et Al. (2005) An educational data mining tool to browse tutor–student interactions: Time will tell! In: Proceedings of the workshop on educational data mining, pp. 15–22.

Okonkwo, R.O. and Enem, F.O. (2011) “Combating Crime and Terrorism Using Data Mining Techniques”, 10th International Conference Information Technology for People-Centered Development (ITePED).

Ricarte, I. L. M., Falci Junior, G. R. (2011) “A Methodology for Mining Data from Computer-Supported Learning Environments”, Informática na Educação: teoria & prática, Porto Alegre, v. 14, n. 2, p. 83-94.

Rogers, Y., Sharp, H., and Preece, J. (2011) Interaction Design: Beyond Human - Computer Interaction. Wiley, 3rd ed.

Romero, C., Ventura, S. and García, E. (2008) “Data mining in course management systems: Moodle case study and tutorial”, Computers & Education 51 (1), 368-384, 405.

Silva, A., et al. (2013) Análise de Redes Sociais para avaliação e monitoramento de programas de treinamento em larga escala baseados no uso de ambientes de aprendizagem e redes sociais online. II Brazilian Workshop on Social Network Analysis and Mining.

Swedan, M.I. (2012) “Students Learning Behavior in Moodle System Using DataMining Techniques”, In: Universal Journal of Applied Computer Science and Technology, 2 (4): 318-823.

Wicker, P. and Breuer, C. (2010) “Analysis of problems using Data Mining techniques – findings from sports clubs in Germany”, In: European Journal for Sport and Society, vol. 7 (2), p.: 131-140

Witten, I.H., Frank, E. and Hall, M. A. (2011) Data Mining: Practical Machine Learning Tools and Techniques, San Francisco: Morgan Kaufmann, 3ed.