Eros Estevão de Moura

Utilização de Técnicas de Mineração de Textos para Apoio a Aprendizagem

Campos dos Goytacazes - RJ

Setembro / 2006

Eros Estevão de Moura

Utilização de Técnicas de Mineração de Textos para Apoio a Aprendizagem

Dissertação apresentada ao Curso de Mestrado em Informática Aplicada da Universidade Cândido Mendes, como requisito parcial para a obtenção do grau de Mestre. Área de Concentração: Sistemas de Informação e Apoio à Decisão.

Orientadora:
Prof$^{a.}$ Dr$^{a.}$ Sahudy Montenegro González

MESTRADO EM INFORMÁTICA APLICADA
UNIVERSIDADE CÂNDIDO MENDES - CAMPOS

Campos dos Goytacazes - RJ

Setembro / 2006

Dissertação de Mestrado sob o título *"Utilização de Técnicas de Mineração de Textos para Apoio a Aprendizagem"*, defendida por Eros Estevão de Moura e _____ em 21 de setembro de 2006, em Campos dos Goytacazes, Rio de Janeiro, pela banca examinadora constituída pelos doutores:

Prof$^{a.}$ Dr$^{a.}$ Sahudy Montenegro González
Departamento de Informática - UCAM-Campos
Orientadora

Prof$^{a.}$ Dr$^{a.}$ Annabell Del Real Tamariz
Departamento de Informática - UCAM-Campos

Prof$^{a.}$ Dr$^{a.}$ Clevi Elena Rapkiewicz
LEPROD / CCT - Universidade Estadual do Norte
Fluminense Darcy Ribeiro-UENF

Dedico esta dissertação a minha família,
cujo apoio em todos os momentos foi fundamental
para o sucesso desse trabalho.

Agradecimentos

Dedico meus sinceros agradecimentos para:

– a professora DSc Sahudy Montenegro González, pela orientação e incentivo;

– o professor DSc Leandro Krug Wives, pela orientação no caminho do Text Mining;

– a professora MSc Jacqueline Uber Silva, pela paciência e ajuda em Text Mining;

– todos os colegas do Mestrado da UCAM-Campos.

"Todo Homem, por natureza, deseja o Conhecimento."

Aristóteles(384-322 A.C.)

Resumo

Na Educação a Distância (EaD), um dos maiores obstáculos para o acompanhamento do aprendizado é o contato entre professor e aluno, já que, na maior parte do tempo, é mediado pelo computador, limitando-se às mensagens escritas. Assim, um das diferenças entre o acompanhamento na educação presencial e na educação a distância está na observação realizada pelo professor, que não pode mais contar com o "corpo a corpo" da sala de aula. Nos ambientes virtuais de aprendizagem, a observação é feita por meio das interações do aluno com o ambiente. Para lidar com essa situação, este trabalho propõe a utilização de técnicas de mineração de textos para apoiar a aprendizagem dos alunos. Os alunos podem expor suas dúvidas em linguagem natural, que serão respondidas com a melhor resposta possível, obtida a partir da mineração da base de conhecimento preenchida pelo professor especialista da disciplina. A interface pode ficar disponível em listas de discussão, bate-papo (*chats*), fóruns, que são mecanismos de comunicação do ambiente de aprendizagem. Para validar as idéias propostas, é apresentado um estudo de caso, desenvolvido dentro do ambiente do Teleduc, os testes feitos e seus resultados.

Palavras-Chave: educação a distância, mineração de texto, mineração de dados

Abstract

One of the largest obstacles in Distance Learning is, literally, the geographical distance between students and teachers and, as a consequence, the lack of face-to-face contact in the learning process. One of the differences between presential and distance learning is the accompanying process by teachers to students in classrooms. In computational environments for distance learning, this supervision is based on student-environment interactions. To deal with this problem, this paper describes a tool to use text mining to support distance learning. Students can expose their doubts in natural language and the tool returns the best feasible answer obtained from the knowledge base data mining. The knowledge base is constructed by specialist teachers of each discipline. The tool interface can be available in forums, chats and discussion lists in the distance learning environment. A case study was developed using the TelEduc environment. Tests and their results measure the proposal effectiveness.

Keywords: Distance Learning, Text Mining, Data Mining

Sumário

Lista de Figuras

Lista de Tabelas

Lista das Abreviaturas

1 Introdução p. 16

2 Tópicos sobre Mineração de Informações p. 19

2.1 Mineração de Dados . . . . . p. 20

2.1.1 Tarefas da Mineração de Dados . . . . . p. 21

Classificação e Regressão . . . . . p. 21

Associação . . . . . p. 21

Segmentação ou *Clustering* . . . . . p. 22

Sumarização e Generalização . . . . . p. 22

2.1.2 Técnicas de Mineração de Dados . . . . . p. 23

2.1.3 A Escolha da Técnica de Mineração de Dados mais adequada . p. 23

2.1.4 Áreas de Aplicação de Técnicas de Mineração de Dados . . . . . p. 26

2.1.5 Estado da Arte em Mineração de Dados . . . . . p. 27

2.1.6 Uma ferramenta para construção da árvore de decisão . . . . . p. 31

Arquivo ARFF do WEKA . . . . . p. 32

Resultados obtidos pelo WEKA . . . . . p. 36

2.2 Mineração de Texto . . . . . p. 39

2.2.1 Conceitos básicos sobre Mineração de Texto . . . . . p. 40
Stopwords . . . . . p. 40
Corpus . . . . . p. 41
Keywords . . . . . p. 42
Collocations . . . . . p. 43
Stemming . . . . . p. 44
Limpeza dos dados (Data Cleaning) . . . . . p. 44
2.2.2 Descoberta reativa e pro-ativa . . . . . p. 45
2.2.3 Tarefas mais comuns de descoberta de conhecimento em textos . p. 45
2.2.4 Recuperação de Informação . . . . . p. 46
2.2.5 Componentes de um Sistema de Recuperação de Informação . . p. 46
2.2.6 Avaliação de sistemas de recuperação de informação . . . . . p. 47
2.2.7 Modelos Clássicos . . . . . p. 48
Modelo Booleano . . . . . p. 48
Modelo Vetorial . . . . . p. 50
Modelo Probabilístico . . . . . p. 51
2.2.8 Relação das principais ferramentas de Mineração de Texto . . . p. 51
TextAnalyst . . . . . p. 51
Market Signal Analyzer . . . . . p. 51
C-4-U Scout . . . . . p. 52
Knowledge Works . . . . . p. 52
Clear Research Suite . . . . . p. 52
BrandPulse . . . . . p. 53
TrackEngine . . . . . p. 53
Strategy . . . . . p. 53
PlanBee . . . . . p. 54

Wincite . . . . . p. 54

Wisdom Builder . . . . . p. 54

S-Miner . . . . . p. 55

Eureka . . . . . p. 55

3 Ensino a Distância p. 56

3.1 Estado da arte na comunicação do aluno em ambientes de EaD . . . . . p. 58

4 Uma proposta para apoiar a aprendizagem em EaD p. 62

4.1 Utilizando mineração de dados . . . . . p. 62

4.1.1 Camada de acesso aos dados . . . . . p. 64

4.1.2 Camada de inteligência . . . . . p. 64

4.1.3 Conclusões da utilização de mineração de dados . . . . . p. 65

4.2 Utilizando mineração de texto . . . . . p. 66

4.2.1 Módulo de manutenção . . . . . p. 68

4.2.1.1 Consulta, alteração e exclusão da base de conhecimento p. 69

4.2.2 Módulo de Inteligência . . . . . p. 69

4.2.3 O módulo de *Interface* . . . . . p. 71

4.3 Problemas e soluções para a ferramenta Text EaD . . . . . p. 73

5 Resultados e Testes p. 74

5.1 O ambiente para a Mineração de Dados . . . . . p. 74

5.1.1 Camada de *interface* . . . . . p. 75

5.1.2 Aplicando a ferramenta . . . . . p. 75

5.1.3 Resultados com a Mineração de Dados . . . . . p. 76

5.2 O ambiente para a Mineração de Texto . . . . . p. 77

5.2.1 Resultados com a Mineração de Texto . . . . . p. 78

6 Conclusões e Trabalhos Futuros p. 79

Referências p. 81

Anexo A – Relação dos códigos fonte da ferramenta Text EaD p. 84

A.1 Programa principal do Text EaD . . . . . p. 84

A.2 Programa para armazenar os documentos em PHP . . . . . p. 85

A.3 Programa para alterar os documentos em PHP . . . . . p. 87

A.4 Programa escolhe o documento em PHP . . . . . p. 89

A.5 Programa mineração de texto em PHP . . . . . p. 90

A.6 Programa retorna os documentos em PHP . . . . . p. 94

A.7 Programa para fazer a pergunta em PHP . . . . . p. 99

Anexo B – Programas que utilizam MD com mecanismo para acesso aos dados p. 101

B.1 Programa principal da ferramenta feito em Java . . . . . p. 101

B.2 Programa gera arquivo WEKA feito em Java . . . . . p. 106

B.3 Programa que processa a MD feito em Java . . . . . p. 108

B.4 Programa que retorna uma resposta feito em PHP . . . . . p. 115

Lista de Figuras

1 Os passos do processo KDD . . . . . p. 20

2 Tela inicial do pacote WEKA . . . . . p. 32

3 Algoritmos implementados pelo WEKA . . . . . p. 33

4 Arquivo no formato ARFF do WEKA . . . . . p. 34

5 Carregando o arquivo ARFF no WEKA . . . . . p. 35

6 Resultado Gerado pelo WEKA . . . . . p. 36

7 Árvore de decisão gerada pelo programa WEKA . . . . . p. 37

8 Os passos do processo Mineração de Texto . . . . . p. 40

9 Lista de *Stopwords* . . . . . p. 41

10 Componentes de um sistema de recuperação de informação . . . . . p. 47

11 Arquitetura da ferramenta . . . . . p. 64

12 Camada de Inteligência da Ferramenta . . . . . p. 65

13 Arquitetura da ferramenta . . . . . p. 68

14 Tela de Cadastro na Base de Conhecimento . . . . . p. 69

15 Página do bate-papo do TelEduc . . . . . p. 71

16 Resposta da ferramenta . . . . . p. 72

17 Página do bate-papo do Teleduc . . . . . p. 75

18 Resposta à dúvida apresentada pelo aluno . . . . . p. 76

Lista de Tabelas

1 Técnicas de mineração de dados . . . . . p. 23

2 Características de dados . . . . . p. 26

3 Características gerais da ferramenta . . . . . p. 28

4 Conectividade a bancos de dados da ferramenta . . . . . p. 29

5 Características de mineração de dados da ferramenta . . . . . p. 30

6 Ferramentas de mineração de dados . . . . . p. 31

7 Modelo de classificação bivalorada . . . . . p. 37

8 Interpretação dos valores de Kappa . . . . . p. 38

9 Classificação das respostas cadastradas . . . . . p. 74

10 Classificação das perguntas realizadas . . . . . p. 77

11 Classificação das áreas de conhecimento . . . . . p. 77

Lista das Abreviaturas

1. AM - Aprendizado de Máquina
2. ARFF - Attribute-Relation File Format
3. AVA - Ambientes Virtuais de Aprendizagem
4. CHAT - Neologismo para designar aplicações de conversação em tempo real
5. DCBD - Descoberta de Conhecimento em Bases de Dados
6. DCT - Descoberta de Conhecimento em Texto
7. DOC - Documento formato Microsoft Word
8. EaD - Educação a Distância
9. EI - Extração de Informação
10. HTML - HyperText Markup Language
11. IC - Inteligência Competitiva
12. IR - Information Retrieval
13. KDD - Knowledge Discovery in Databases
14. MD - Mineração de Dados
15. MT - Mineração de Textos
16. PDF - Portable Document Format
17. PHP - PHP: Hypertext Preprocessor
18. RI - Recuperação da Informação
19. RTF - Rich Text Format
20. SGBD - Sistema Gerenciador de Banco de Dados

21. SQL - Structured Query Language

22. SRI - Sistema de Recuperação de Informação

23. TELEDUC - Ambiente de Ensino a Distância - UNICAMP

24. TIC - Tecnologias de Informação e Comunicação

25. UNIFOR - Universidade de Fortaleza

26. URL - Uniform Resource Locator

27. WEB - World Wide Web

28. WEKA - Waikato Environment for Knowledge Analysis

1 Introdução

Ao longo do tempo, o desenvolvimento de tecnologias associadas à informação e comunicação tem sido um agente relevante na aprendizagem. A Educação a Distância (EaD) é uma destas evoluções, que associa informação e comunicação com novas tecnologias.

Um dos grandes desafios em educação a distância, hoje, é como superar as dificuldades impostas pela própria distância. Assim sendo, o acompanhamento da aprendizagem na EaD pode tirar proveito das inúmeras vantagens criadas pelas tecnologias, mas para isto, precisa-se criar modelos que, efetivamente, contenham alguma mudança qualitativa que ajudem a diminuir a distância (ou aumentar a comunicação) entre o aluno e o professor.

As novas tecnologias de comunicação, que estão sendo colocadas à disposição dos alunos e dos centros produtores, têm facilitado muito a ligação do aluno aos apoios didáticos, pela rapidez e pelos baixos custos. Estas mesmas tecnologias estão possibilitando um salto de qualidade na comunicação, produzindo mecanismos de contato entre os alunos, mesmo a distância, para que troquem experiências e vivências. Um dos meios mais apropriados para tal, dado o baixo custo, é o correio eletrônico.

Com o desenvolvimento da Web, surgiram os ambientes de EaD. As fronteiras para a educação a distância se expandiram, podendo reunir-se em um só meio de comunicação as vantagens dos diferentes modos de se comunicar informações e idéias, de forma cada vez mais interativa, reduzindo-se custos e ampliando as possibilidades de auto-descobrimento, através do uso de milhares de opções de buscas de informações na grande rede mundial e mecanismos de comunicação como os *chats*, *e-mails* e conferências eletrônicas.

Em (1) foi realizado um levantamento sobre alguns ambientes para EaD, com o objetivo de identificar quais recursos os mesmos disponibilizam para o acompanhamento do aprendizado do aluno. A partir disso, foi elaborada uma tabela, que traz uma relação de 18 ambientes de EaD, seus mecanismos de avaliação e acompanhamento e a freqüência desses mecanismos nos ambientes pesquisados. Pode-se constatar que esses ambientes de EaD tendem a usar provas *on-line* e rastrear algumas atividades dos alunos para acom-

panhar o aprendizado. 94% dos ambientes fornecem mecanismos para avaliação formal, através de provas, e 78% dos mesmos ambientes oferecem recursos para monitoramento das atividades dos alunos em forma de rastreamento das ações dos mesmos, sendo que apenas menos dos 20% dos ambientes usam mecanismos de comunicação, como listas de discussão, *newsgroups*, fórum, *e-mails* e bate-papo para interagir com os alunos.

Considerando esta problemática, este artigo propõe aumentar a interação do aluno com os ambientes de EaD, aproveitando os mecanismos de comunicação listados acima. A idéia é que a introdução de técnicas de mineração de texto (*text mining*) no ambiente permitirá ao aluno tirar maior proveito e melhor se comunicar com o ambiente. Assim sendo, técnicas efetivas tornarão o ambiente de aprendizagem mais confiável e os alunos passarão a utilizá-lo com maior freqüência.

O uso de mineração de texto é justificado segundo vários motivos: (1) à medida que um curso à distância é concebido para atender ao público remoto, o número de alunos pode crescer consideravelmente, e, portanto, a tarefa de acompanhar o estudante pode-se tornar difícil; (2) normalmente, os cursos de EaD mantêm seus dados em bancos de dados e a natureza histórica destes dados pode ser útil para análises prospectivas; (3) os dados históricos podem servir para encontrar padrões na comunicação do aluno com o ambiente de ensino; (4) (2) destaca que a implantação de um programa de coleta e análise de dados pode levar a melhorias na educação como nenhuma outra inovação o fez; (5) pouco se tem feito em termos de sistemas de suporte a decisão em EaD. A maioria dos ambientes existentes não oferece recursos sofisticados para apoiar decisões.

Nesse contexto, propor ações que atuem sobre pelo menos alguns dos fatores apontados torna-se iminente. O foco atendido por este trabalho é o de integrar a busca e descoberta de conhecimento que auxiliem ao aluno em um ambiente Web de educação a distância, para apoiar o processo de ensino/aprendizagem dos alunos e promover a comunicação aluno-ambiente. Para isto, optou-se pelo desenvolvimento de uma ferramenta cuja finalidade é a de oferecer aos alunos suporte para que possam expor, em linguagem natural (português), suas dúvidas nas diferentes disciplinas do curso de EaD. As perguntas serão respondidas com a melhor resposta possível, a partir da utilização de técnicas para minerar texto em bases de conhecimento, preenchidas pelos professores especialistas das disciplinas. A interface da ferramenta no ambiente de aprendizagem pode ficar disponível em qualquer um dos mecanismos de comunicação.

Para apresentar tal proposta, este trabalho está organizado em seis capítulos. No capítulo 2, é realizado um estudo sobre mineração. No capítulo 3, são relatados os meca-

nismos de acompanhamento de alunos em ambientes de EaD. No capítulo 4, é especificada a ferramenta, sua modelagem e implementação. No capitulo 5, é descrito o estudo de caso. Por último, o capítulo 6 apresenta as conclusões e trabalhos futuros.

Como parte deste trabalho, os autores desenvolveram três artigos, citados a seguir.

- Uma Ferramenta para o Apoio ao Aprendizado em um Ambiente de Educação a Distância - Aceito no Simpósio Mineiro de Sistemas de Informação - SMSI 2006 - SBC.
- Uma Ferramenta de Mineração de Dados para Apoio ao Aprendizado em um Ambiente de Educação a Distância - Aceito no V Simpósio de Informática da Região Centro do RS, 2006 - SIRC/RS 2006 - SBC
- Uma Proposta para Apoiar o Aprendizado em Ambientes de Educação a Distância com a Utilização de Mineração de Texto - Submetido no XVII Simpósio Brasileiro de Informática na Educação - SBIE 2006 - SBC

2 *Tópicos sobre Mineração de Informações*

No mundo contemporâneo, devido a crescente participação dos computadores na sociedade, há uma grande quantidade de informação sendo gerada. Seja na realização de uma chamada telefônica ou na utilização de um cartão de crédito, temos computadores alimentados repetidamente por sistemas de monitoramento e sensores. Nos negócios, engenharia, ciência, medicina, nas instituições governamentais e comerciais, há uma enorme quantidade de máquinas prontas para capturar todo o conhecimento gerado nas diversas atividades realizadas. No entanto, esta grande quantidade de informação é ainda subexplorada, à medida que decisões são tomadas, a todo instante, sem levar em conta essa riqueza de informações. Conseqüentemente, tais decisões podem ser não ótimas ou mesmo erradas (**??**). Atualmente, no mundo moderno e globalizado, as empresas buscam absorver o conhecimento de forma rápida e segura.

Para que uma parcela maior das informações reunidas alcance a finalidade para a qual foram coletadas, o homem conta com o poder oferecido pelos sistemas de *hardware* computacionais, de natureza digital, adequados ao processamento maciço de informação. No entanto, para que esta facilidade possa ser aproveitada são necessários *softwares* capazes de promover a investigação dos dados armazenados. Porém, o desenvolvimento destes *softwares*, até então, ainda não é capaz de decifrar grandes quantidades de dados, pois carecem de meios mais poderosos de investigação.

As ferramentas de exploração de dados a serem desenvolvidas devem buscar escalabilidade e poder investigativo, este último só podendo ser alcançado através de engenhosas interfaces de interação com o homem, pois se sabe que o processo de descoberta não pode ser totalmente automatizado (**??**) já que engloba inteligência e criatividade, características que o computador ainda não é capaz de simular.

O homem ainda irá atuar decisivamente na utilização destes sistemas, que devem auxiliá-lo adequadamente (**??**). Nesta perspectiva se encaixa a especialidade da ciência de

computação denominada *Knowledge Discovery in Databases* (KDD). O processo do KDD pode ser observado na Figura **??**. Um processo complexo que objetiva extrair conhecimento a partir de grandes volumes de dados. O KDD é um processo de investigação constituído por várias etapas: seleção, pré-processamento, transformação, Mineração de Dados (MD) e interpretação/avaliação (**??**). Sua demanda vem impulsionando, principalmente, as pesquisas por novas técnicas de MD, que é o núcleo de todo processo. Essa tecnologia não está disponível para a maioria das empresas, seja por custo, falta de conhecimento sobre a MD ou por falta de técnicos preparador para este tecnologia. Segundo (**??**) a MD é uma das etapas do KDD.

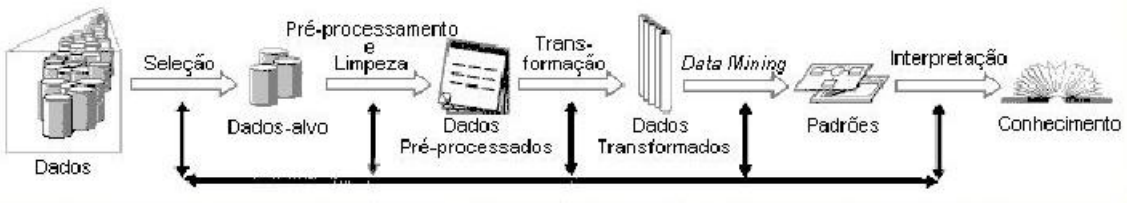

Figura 1: Os passos do processo KDD

2.1 Mineração de Dados

A construção de modelos estatísticos costuma ser o método tradicionalmente utilizado para derivar tendências. Entretanto, estatísticas tradicionais são limitadas se considerarmos que:

- a análise se torna trabalhosa quando o número de variáveis a serem investigadas cresce;
- seguidamente, métodos estatísticos possuem condições que limitam o número de casos a utilizar, fazendo com que apenas uma pequena parte do universo esteja disponível para a análise;
- quando os relacionamentos dos dados são não lineares, torna-se difícil aplicar métodos estatísticos tradicionais

Além disso, poucas empresas dispõem de pessoal preparado para esta tarefa tão especializada, mesmo com a ajuda de estatísticos, pode-se levar semanas para projetar e construir os modelos.

Apesar do termo mineração de dados (MD) ter se tornado bastante popular nos últimos anos, existe ainda certa confusão quanto à sua definição. MD é o processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases de dados (**??**). Segundo (**??**) *data mining* é definido como uso de técnicas automáticas de exploração de grandes volumes de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano.

2.1.1 Tarefas da Mineração de Dados

Podemos observar para quais tarefas a MD pode ser utilizada nos itens a seguir :

Classificação e Regressão

Classificação e regressão usam dados existentes para criar modelos de comportamento de variáveis. A operação de classificação cria automaticamente um modelo a partir de um conjunto inicial de registros. Esse conjunto serve de exemplo e é chamado de conjunto de treinamento. Os registros do conjunto de treinamento devem pertencer a um pequeno grupo de classes predefinidas. O modelo é composto de padrões, essencialmente generalizações em relação aos registros, os quais são usados para diferenciar as classes. Uma vez obtido o modelo, este é usado para classificar automaticamente os demais registros.

O modo como às classes são criadas oferece vantagens em relação a métodos estatísticos. Os padrões podem ser produzidos a partir de um conjunto localizado de fenômenos, ao passo que métodos estatísticos devem agir sobre populações inteiras e de distribuição bem conhecida. Desta forma é possível prever características de um pequeno percentual do conjunto de registros, o que não seria alcançado estatisticamente dado à inexpressividade dos registros sendo avaliados. Como exemplo, uma empresa de cartão de crédito poderia examinar algumas características de seus clientes e prever o nível de inadimplência associado. Tais características poderiam incluir renda, histórico de crédito, tipo e localização do emprego.

Associação

Associações são relacionamentos significativos entre itens de dados armazenados. O objetivo da operação é encontrar tendências, a partir de grande número de transações, que possam ser usadas para entender e explorar padrões de comportamento dos dados.

Um exemplo seria o de varrer registros de terminais de ponto de venda e descobrir que tipos de itens são vendidos juntos, de forma a redefinir a disposição dos artigos na loja e sua promoção em campanhas publicitárias, permitindo explorar com maior eficácia essas associações.

Segmentação ou *Clustering*

O agrupamento em *clusters* envolve segmentar a informação disponível em conjuntos definidos e homogêneos baseando-se em atributos específicos. O conceito de *clustering* já tem uma longa história em estatística, mas o que tem de novo em MD é o fato de poder também ser aplicada a itens não numéricos. Os resultados de uma operação de clusterização podem ser usados de duas diferentes maneiras: para produzir um sumário da base de dados ou como dados de entrada para outras técnicas, por exemplo, classificação, já que um *cluster* é um grupo menor e de mais fácil manuseio por parte de algoritmos de classificação.

Especialmente devido ao alto custo envolvido, estas ferramentas vinham sendo usadas, até o momento, quase que unicamente por grandes corporações e instituições governamentais. A maior parte das atividades de MD ficava restrita a especialistas, com empresas oferecendo seus serviços de análise, mas sem entregar aos clientes seus métodos e ferramentas. Com o grande aumento do volume de dados nas empresas e com o crescimento do uso de tecnologia de banco de dados, especialmente de *data warehouse*, as técnicas de MD assumiram papel importante no suporte aos processos de tomada de decisão e devem, aos poucos, ganhar mercado dentre empresas de menor porte. No entanto, no atual estado da arte destas ferramentas, ainda requerem um bom nível de conhecimentos do domínio da aplicação, de estatística e da própria ferramenta. No livro Intranet *Data Warehouse* (**??**) afirma que não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e suas desvantagens sendo mais específico para um problema.

Sumarização e Generalização

Como o próprio nome diz, esta tarefa procura gerar uma caracterização de um conjunto de dados fornecidos. Por exemplo, a partir de um banco de dados de um supermercado, poder-se-ia caracterizar que os clientes que compram cerveja e carne de churrasco são casados, com mais de 30 anos e pertencem a uma determinada faixa salarial;

2.1.2 Técnicas de Mineração de Dados

Harrison (**??**) afirma que não há uma técnica que resolva todos os problemas de MD. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados. A Tabela **??** apresenta um resumo das técnicas de mineração de dados normalmente usadas.

Técnicas	Descrição	Tarefas
Descoberta de regras de associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados	Associação.
Árvores de decisão	Hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos	Classificação e regressão
Raciocínio	baseado em Casos. Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança	Classificação, segmentação
Algoritmos genéticos	Métodos gerais de busca e otimização, inspirados na teoria da evolução, onde a cada nova geração, soluções melhores têm mais chance de ter "descendentes"	Classificação, segmentação
Redes neurais artificiais	Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões	Classificação, segmentação

Tabela 1: Técnicas de mineração de dados

2.1.3 A Escolha da Técnica de Mineração de Dados mais adequada

A escolha de uma técnica de MD a ser aplicada não é uma tarefa fácil. Segundo Harrison (**??**), a escolha das técnicas de MD dependerá da tarefa específica a ser executada e dos dados disponíveis para análise. Berry e Linoff (**??**) sugerem que a seleção das técnicas de MD deve ser dividida em dois passos:

1. traduzir o problema de negócio a ser resolvido em séries de tarefas de MD;
2. compreender a natureza dos dados disponíveis em termos de conteúdo, tipos de campos de dados e a estrutura das relações entre os registros.

O primeiro passo na seleção da técnica de MD é, portanto, estabelecer uma meta comercial como, por exemplo, a estratégia de identificar assinantes que tenham a intenção de desistir do serviço, descobrir suas razões para isso e fazer algum tipo de oferta especial que os agrade. Para o sucesso da estratégia, é preciso não somente identificar os assinantes que podem cancelar, mas dividi-los em grupos de acordo com seus motivos presumíveis para a desistência. A primeira tarefa é, obviamente, a classificação. Usando um conjunto de dados de treinamento com exemplos de clientes que cancelaram o serviço juntamente com exemplos daqueles que permaneceram, é possível construir um modelo capaz de rotular cada cliente como "fiel" ou "instável".

O segundo passo seria determinar as características dos dados em análise. A meta é selecionar a técnica de MD que minimiza o número e dificuldades de transformação de dados para, a partir destes, obter bons resultados.

A Tabela **??** mostra uma lista de características de dados, baseada em Berry e Linoff (**??**), que ajudará na escolha de uma abordagem da MD.

Característica	Descrição	Técnicas de mineração de dados
Variáveis de categorias	São campos que apresentam valores de um conjunto de possibilidades limitado e predeterminado	Descoberta de regras de associação árvores de decisão
Variáveis numéricas	São aquelas que podem ser somadas e ordenadas	Raciocínio baseado em casos ou árvores de decisão

Continua na próxima página

Característica	Descrição	Técnicas de mineração de dados
Muitos campos por registros	Este pode ser um fator de decisão da técnica correta para uma aplicação específica, uma vez que os métodos de mineração de dados variam na capacidade de processar grandes números de campos de entrada	árvores de decisão
Variáveis dependentes múltiplas	Caso em que é desejado prever várias variáveis diferentes baseadas nos mesmos dados de entrada	Redes neurais
Registro de comprimento variável	Apresentam dificuldades na maioria das técnicas de mineração de dados, mas existem situações em que a transformação para registros de comprimento fixo não é desejada	descoberta de regras de associação
Dados ordenados cronologicamente	Apresentam dificuldades para todas as técnicas e, geralmente, requerem aumento dos dados de teste com marcas ou avisos, variáveis de diferença etc.	Rede neural intervalar {time-delay} ou Descoberta de regras de associação

Continua na próxima página

Característica	Descrição	Técnicas de mineração de dados
Texto sem formatação	A maioria das técnicas de mineração de dados é incapaz de manipular texto sem formação	Raciocínio baseado em casos

Tabela 2: Características de dados

2.1.4 Áreas de Aplicação de Técnicas de Mineração de Dados

A seguir, são relacionadas as principais áreas de interesse na utilização de mineração de dados:

- *Marketing*: Técnicas de MD são aplicadas para descobrir preferências do consumidor e padrões de compra, com o objetivo de realizar {marketing} direto de produtos e ofertas promocionais, de acordo com o perfil do consumidor.
- Detecção de fraudes: muitas fraudes óbvias (tais como, a compensação de cheque por pessoas falecidas) podem ser encontradas sem MD, mas padrões mais sutis de fraude podem ser difíceis de serem detectados, por exemplo, o desenvolvimento de modelos que predizem quem será um bom cliente ou aquele que poderá se tornar inadimplente em seus pagamentos.
- Medicina: caracterizar comportamento de paciente para prever visitas, identificar terapias médicas de sucesso para diferentes doenças, buscar por padrões de novas doenças.
- Instituições governamentais: descoberta de padrões para melhorar as coletas de taxas ou descobrir fraudes.
- Ciência: técnicas de MD podem ajudar cientistas em suas pesquisas, por exemplo, encontrar padrões em estruturas moleculares, dados genéticos, mudanças globais de clima, oferecendo conclusões valiosas rapidamente.
- Controle de processos e controle de qualidade: auxiliar no planejamento estratégico de linhas de produção e buscar por padrões de condições físicas na embalagem e armazenamento de produtos.

- Banco: detectar padrões de uso de cartão de crédito fraudulento, identificar clientes "ilegais", determinar gastos com cartão de crédito por grupos de clientes, encontrar correlações escondidas entre diferentes indicadores financeiros.

- Apólice +de seguro: análise de reivindicações determinar quais procedimentos médicos são solicitados em conjunto, prever quais clientes comprarão novas apólices, identificar padrões de comportamento de clientes perigosos, identificar comportamento fraudulento.

- Transporte: determinar as escalas de distribuição entre distribuidores, analisar padrões de carga.

- C & T (Ciência e Tecnologia): avaliar grupos de pesquisa do país (Gonçalves, 2000), (Romão, 1999), (Dias, 2001).

- Web: existem muitas pesquisas direcionadas à aplicação de MD na Web, tais como: (Loh et al, 2000), (Kosala e Blockeel, 2000), (Ma et al, 2000), (Mobasher et al, 2000), (Sarawagi e Nagaralu, 2000), (Spiliopoulou, 2000).

2.1.5 Estado da Arte em Mineração de Dados

De acordo com Goebel e Gruenwald (**??**), muitas ferramentas atualmente disponíveis são ferramentas genéricas da Inteligência Artificial ou da comunidade de estatística. Tais ferramentas geralmente operam separadamente da fonte de dados, requerendo uma quantidade significativa de tempo gasto com exportação e importação de dados, pré e pós processamento e transformação de dados.

Para Goebel e Gruenwald, as características a serem consideradas na escolha de uma ferramenta de descoberta de conhecimento devem ser as seguintes:

- a habilidade de acesso a uma variedade de fontes de dados, de forma on-line e offline;

- a capacidade de incluir modelos de dados orientados a objetos ou modelos não padronizados (tal como multimídia, espacial ou temporal);

- a capacidade de processamento com relação ao número máximo de tabelas/tuplas/atributos;

- a capacidade de processamento com relação ao tamanho do banco de dados;

- variedade de tipos de atributos que a ferramenta pode manipular;

- tipo de linguagem de consulta.

Goebel e Gruenwald propõem, também, um esquema de classificação de características que pode ser usado para estudar ferramentas de descoberta de conhecimento e de mineração de dados. Neste esquema, as características das ferramentas são classificadas em três grupos: características gerais, conectividade a banco de dados e características de MD.

As Tabelas **??**, **??** e **??** mostram como as características das ferramentas são classificadas de acordo com esses grupos.

Característica	Classificação
Produto	Nome e vendedor do produto de software
Status da Produção	P=Comercial, A=Alfa, B=Beta, R=Protótipo de Pesquisa
Status Legal	PD=Domínio Público, F=Freeware, S=Shareware
Licença Acadêmica	Se existe licença acadêmica livre disponível ou redução de custo
Demo	D=Versão Demo disponível para download na internet, R=Demo disponível através de requisição, U=Não conhecido
Arquitetura	S=Standalone, C/S=Cliente/Servidor, P=Processamento Paralelo
Sistemas Operacionais	Lista de sistemas operacionais para os quais a versão atual do software pode ser obtida.

Tabela 3: Características gerais da ferramenta

Característica	Classificação
Fontes de dados	T=Arquivos texto Ascii, D=Arquivos Dbase,P=Arquivos Paradox, F=Arquivos Foxpro, Ix=Informix, O=Oracle, Sy=Sybase, Ig=Ingres, A=MS Access, OC=Conexão aberta de banco de dados (ODBC), SS=Servidor MS SQL, Ex=MS Excel, L=Lótus 1-2-3.
Conexão a BD	Onl=Online, Offl=Offline
Tamanho	S=Pequeno (até 10.000 registros), M=Mediano (10.000 a 1.000.000 registros), L=Grande (mais de 1.000.000)
Modelo	R=Relacional, O=Orientado a Objetos, 1= Uma Tabela
Atributos	Co=Contínuo, Ca=Categórico (valores numéricos discretos), S=Simbólico
Consulta	S=Linguagem de consulta estruturada (SQL ou derivada), Sp=Uma linguagem de consulta específica, G=Interface gráfica de usuário, N=Não aplicável, U=Não conhecido

Tabela 4: Conectividade a bancos de dados da ferramenta

Característica	Classificação
Fontes de dados	T=Arquivos texto Ascii, DĀrquivos Dbase, P=Arquivos Paradox, F=Arquivos Foxpro, IxĪnformix, OŌracle, SyS̄ybase, IgĪngres, A=MS Access, OCC̄onexão aberta de banco de dados (ODBC), SSS̄ervidor MS SQL, ExM̄S Excel, LL̄ótus 1-2-3.

Tabela 5: Características de mineração de dados da ferramenta

Existem ferramentas que implementam uma ou mais técnicas de MD. A Tabela **??** relaciona algumas dessas ferramentas, fornecendo informações tais como: a empresa fornecedora, as técnicas implementadas de mineração de dados e exemplos de aplicações.

Ferramenta	Empresa Fornecedora	Técnicas de mineração de dados
Aplicações AIRA/ Hycones IT (1998)	Regras de associação	Gerenciamento de relacionamento de cliente, marketing, detecção de fraude, controle de processo e controle de qualidade.
Alice 5.1 & Isoft AS (1998)	Árvore de decisão, raciocínio baseado em casos	Política de crédito, marketing, saúde, controle de qualidade, recursos humanos.
Clementine/ Integral Solutions Limited (ISL, 1996)	Indução de regras, árvores de decisão, redes neurais	O *marketing* direto, identificação de oportunidades de venda cruzada, retenção de cliente, previsão de lucro do cliente, detecção de fraude, segmentação e lucro do cliente.
DataMind/ DataMind Technology Center (1998), (Groth, 1998)	(abordagem própria)	Não identificadas.
Decision Series/ Neovista Solutions Inc. (1998)	Árvore de decisão, métodos estatísticos, indução de regras, redes neurais	Marketing direcionado, detecção de fraude, retenção de cliente, análise de risco, segmentação de cliente, análise de promoção.
Intelligent Miner/ IBM (1997)	Árvores de decisão, redes neurais	Segmentação de cliente, análise de conjunto de itens, detecção de fraude.

Continua na próxima página

Ferramenta	Empresa Fornecedora	Técnicas de mineração de dados
KnowledgeSEEKER / Angoss IL (Groth, 1998)	Árvores de decisão	Indução de regras de lucro e segmentação de cliente para detecção de fraude e análise de risco, controle de processo, marketing direto.
NeuralWorks Predict/ NeuralWare (Groth, 1998)	Rede neural	Indústria.
MineSet/ Silicon Graphics Computer Systems (2000)	Métodos estatísticos, árvores de decisão, indução de regras	Áreas da saúde, farmacêutica, biotecnologia e química.
PolyAnalyst/ Megaputer Intelligence Ltd. (1998)	Algoritmo genético	Métodos estatísticos, indução de regras, marketing direto, pesquisa médica, análise de conjunto de itens.
WEKA (free)	Regras de associação, árvore de decisão	Indução de regras, marketing direto, detecção de fraude, etc.

Tabela 6: Ferramentas de mineração de dados

2.1.6 Uma ferramenta para construção da árvore de decisão

Como pode ser observado em **??** existem algumas ferramentas que implementam árvore de decisão. Será detalhada uma em particula, Weka, por ser desenvolvida na linguagem Java e poder rodar nas mais variadas plataformas (Windows, Linux, Mac Os, etc). Além disso, é um software de domínio público podendo ser obtido em http://www.cs.waikato.ac.nz/ml

Segundo WEKA (2005), o pacote WEKA (Waikato Environment for Knowledge Analysis) é formado por um conjunto de implementações de algoritmos de diversas técnicas de mineração de dados.

Para a implementação da versão da ferramenta Text EaD com mineração de dados foram utilizadas algumas classes da ferramenta WEKA.

Ao se instalar o WEKA, será apresentada uma tela inicial com quatro botões, Figura **??**.

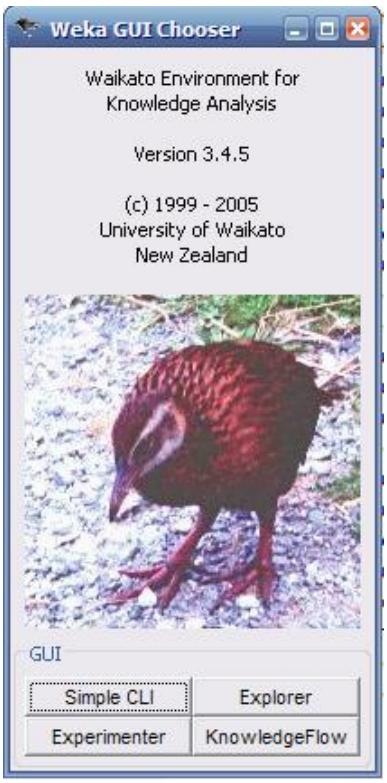

Figura 2: Tela inicial do pacote WEKA

O WEKA implementa vários métodos de classificação: Árvore de Decisão, Regras de Aprendizagem, Naive Bayes, Tabelas de Decisão, Regressão Lógica, entre outros.

Os algoritmos disponíveis para geração de árvore de decisão, podem ser visualizados na Figura **??**:

O WEKA possui um formato próprio para o arquivo de entrada de dados, o ARFF (*Attribute-Relation File Format*). Antes de aplicar os dados a qualquer algoritmo do pacote WEKA estes devem ser convertidos para o formato ARFF.

Arquivo ARFF do WEKA

O formato ARFF consiste basicamente de duas partes. A primeira contém uma lista de todos os atributos $Real, Integer, String, etc$, definidos por um tipo ou por um conjunto de valores. Se utilizarmos, os valores estes devem estar entre "{ }" separados por vírgula. A segunda parte consiste das instâncias, ou seja, os registros a serem minerados com o valor dos atributos para cada instância separada por vírgula, a ausência de um item em um registro deve ser atribuída pelo símbolo de interrogação

"?"

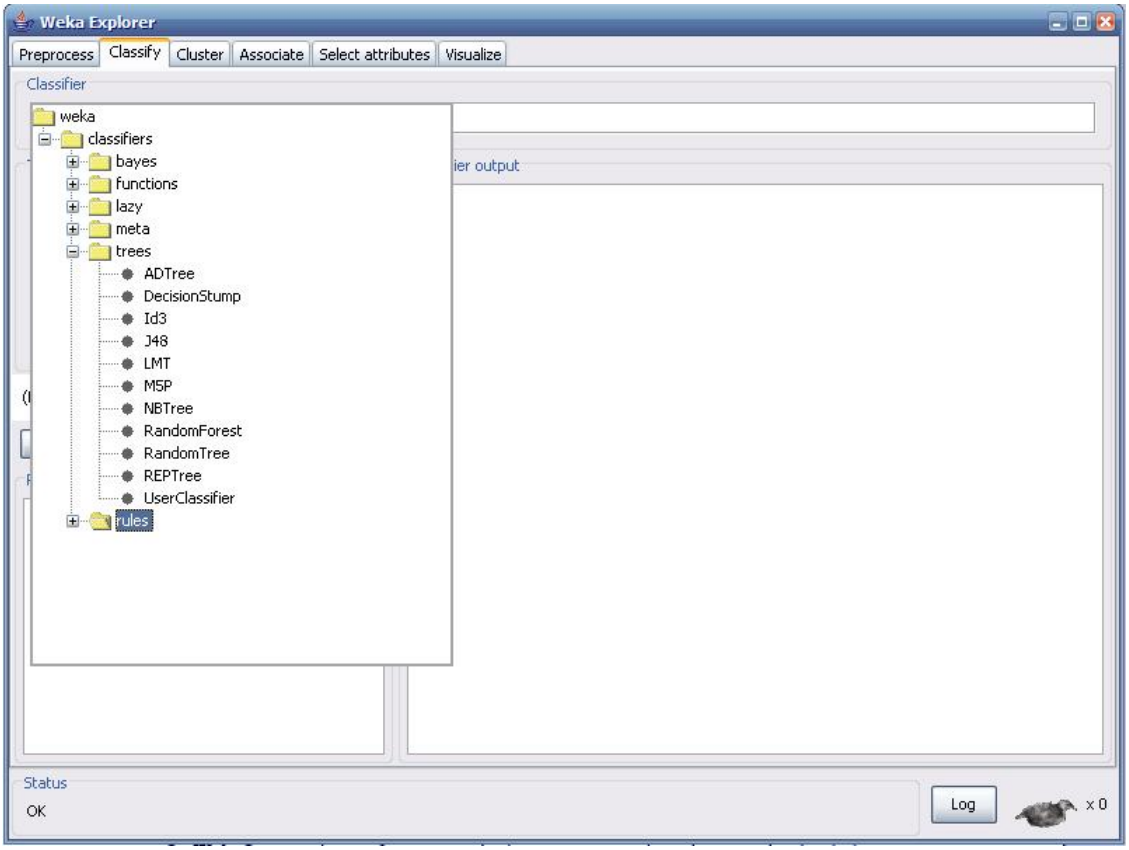

Figura 3: Algoritmos implementados pelo WEKA

Na primeira linha, deve conter o comando relation *nomedoconjuntodedados*, em seguida tem-se a relação dos atributos, onde se coloca o nome do atributo e tipo ou seus possíveis valores, definido por attribute nome_do_atributo tipo ou {valores}.

Após isso deve conter o comando data e nas próximas linhas devem ser listados os registros onde cada linha representa um registro.

O exemplo, na Figura **??**, é instalado junto com o pacote WEKA. O arquivo contém dados atmosférico tais como: temperatura, umidade, vento e como estava o céu. Esses dados foram coletados durante 14 partidas de golfe, e servirão para gerar regras, definindo se deve ou não jogar golfe.

A variável objetivo é joga, podendo apresentar os seguintes resultados: sim ou não.

Para iniciar o WEKA pressiona-se o botão Explorer, a janela WEKA Knowledge Explorer será aberta, deve-se então carregar os dados para serem analisados os quais podem ser originados de um arquivo (Open file...) de uma Uniform Resource Locator (URL) (Open URL...) ou ainda de um banco de dados (Open DB...). Neste exemplo os dados encontram-se em um arquivo ARFF, clicando em Open File carrega-se o arquivo

```
@relation clima

@attribute ceu {sol, nublado, chuva}
@attribute temperatura real
@attribute umidade real
@attribute vento {verdadeiro, falso}
@attribute joga {sim, nao}

@data
sol,85,85,falso,nao
sol,80,90,verdadeiro,nao
nublado,83,86,falso,sim
chuva,70,96,falso,sim
chuva,68,80,falso,sim
chuva,65,70,verdadeiro,nao
nublado,64,65,verdadeiro,sim
sol,72,95,falso,nao
sol,69,70,falso,sim
chuva,75,80,falso,sim
sol,75,70,verdadeiro,sim
nublado,72,90,verdadeiro,sim
nublado,81,75,falso,sim
chuva,71,91,verdadeiro,nao
```

Figura 4: Arquivo no formato ARFF do WEKA

de dados, conforme ilustra a Figura **??**

Na parte superior encontra-se as seguintes abas: Preprocess, onde se pode editar e salvar a base; Classify, conjunto de algoritmos que implementam os esquemas de aprendizagem que funcionam como classificadores; Cluster, contém os algoritmos para geração de grupos; Associate, conjunto de algoritmos para gerar regras de associação, Select attributes determina a relevância dos atributos e Visualise explora os dados.

Em Test options definem-se algumas opções de teste como conjunto de treinamento (Use training set), fornecer um conjunto de teste (Supplied test set), validação cruzada (Cross-validation) com o número de partições e porcentagem dos dados usados para treinamento (Percentage split) em More options... temos algumas opções de saída.

No exemplo, como há poucos registros, utilizou-se os dados como um conjunto de treinamento ativando a opção Use training set.

Selecionou-se o algoritmo J48[1] e clicou-se em Start, o WEKA gerou o seguinte resultado que pode ser observado na figura **??**

[1]O algoritmo J48 é uma implementação do algoritmo C4.5, ambos utilizados para a mineração de documentos textuais. Ele constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, e usa esse modelo para classificar outras instâncias num conjunto de testes.

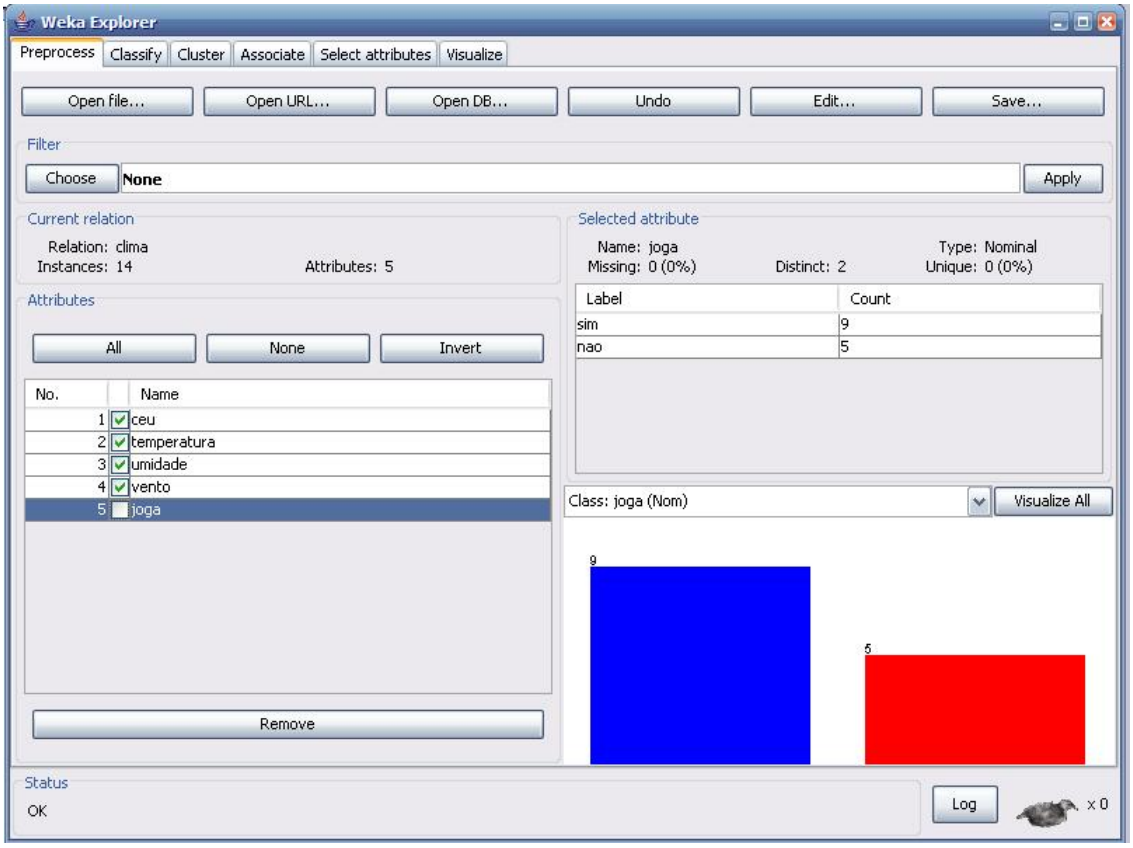

Figura 5: Carregando o arquivo ARFF no WEKA

Na Figura **??** ilustra-se a árvore de decisão seguindo as regras geradas pelo WEKA

Com base na árvore de decisão apresentada na Figura 10, pode-se extrair as seguintes regras:

1. Se ceu = sol e umidade ¿ 75 então jogar = não;
2. Se ceu = sol e umidade ¡= 75 então jogar = sim;
3. Se ceu = nublado então jogar = sim;
4. Se ceu = chuva e vento = verdadeiro então jogar = sim;
5. Se ceu = chuva e vento = falso então jogar = não.

Conclui-se que em 2 situações não será permitido jogar:

1. toda vez que estiver fazendo sol e a umidade relativa do ar for menor que 75
2. se estiver chovendo e não tiver vento.

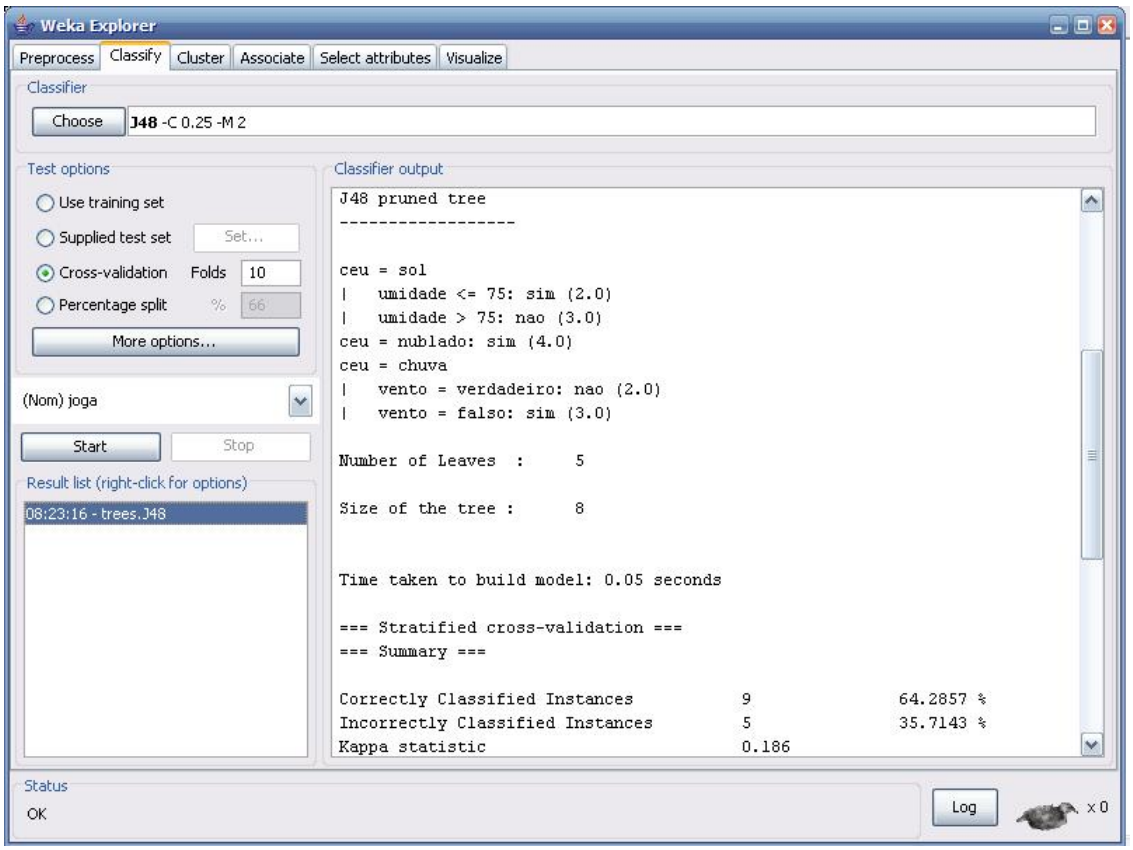

Figura 6: Resultado Gerado pelo WEKA

Várias interpretações podem ser obtidas pelos dados apresentados como resultado, por exemplo, a taxa de erro no caso de permissão para jogar consiste em 77,78% e para não jogar é de 40%.

Resultados obtidos pelo WEKA

A avaliação de um modelo é realizada sobre o conjunto de teste. Em aprendizado de máquina, os métodos usados para seleção dos exemplos são hold-out e crossvalidation, sendo que ambos estimam a qualidade média do modelo construído por meio de várias amostragens. A seguir algumas definições:

Hold-out: um conjunto de teste de tamanho pré-definido é selecionado aleatoriamente, sendo que o restante dos exemplos do conjunto é usado para treinamento do modelo.

Cross-validation: o conjunto de dados é particionado em N subconjuntos sendo que as fases de teste são realizadas em N vezes. Em cada interação, um desses subconjuntos é selecionado para ser o conjunto de teste, enquanto os demais compõem o conjunto de treinamento.

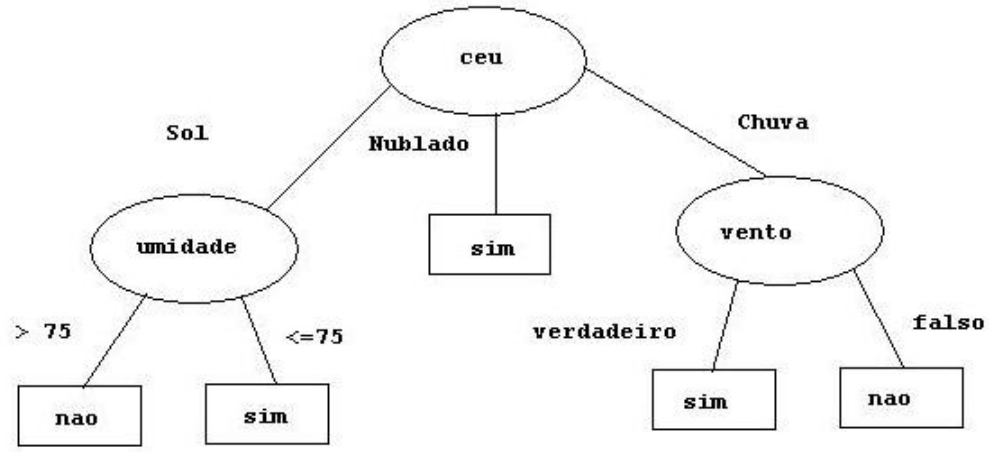

Figura 7: Árvore de decisão gerada pelo programa WEKA

Tabela 7: Modelo de classificação bivalorada

Exemplos	Predito como positivo	Predito como negativo
Exemplos positivos	tp	fp
Exemplos negativos	fn	tn

Na classificação que abrange classes de valores discretos, as estimativas de qualidade podem ser determinadas por meio de uma matriz de confusão obtida a partir dos resultados da classificação. A matriz de confusão envolve os valores reais dos exemplos e os valores preditos pelo modelo. Na Tabela **??** é apresentado um modelo de classificação bi-valorada, na qual tp representa a quantidade de exemplos da classe positiva que foram classificados como positivos (positivo verdadeiro), fp representa a quantidade de exemplos da classe negativa que foram classificados como positivos (falso positivo), fn representa a quantidade de exemplos da classe positiva que foram classificados como negativos (falso negativo), e tn representa a quantidade de exemplos da classe negativa que foram classificados como negativos (negativo verdadeiro).

A partir da matriz de referência cruzada, pode-se obter vários resultados para medir a qualidade da classificação dos textos.

Esses resultados são obtidos das seguintes medidas de avaliação:

Precision, Recall, F-measure e Kappa, que são definidos abaixo:

Precision: a precisão de um modelo é a proporção de exemplos positivos que foram corretamente classificados. Pode ser calculada de acordo com a equação abaixo:

Precision = tp (tp + fp)

Tabela 8: Interpretação dos valores de Kappa

Valores de Kappa	Interpretação
0	Nenhuma concordância
0 0.19	Baixa concordância
0.20 0.39	Concordância médiabaixa
0.40 0.59	Concordância moderada
0.60 0.79	Concordância significativa
0.80 1.00	Concordância quase perfeita

Onde: tp = positivo verdadeiro fp = falso positivo

Recall: é definida como a porção classificada corretamente como exemplos positivos. A estimativa dessa medida é calculada conforme a seguinte a equação:

Recall = tp / (tp + fn)

Onde: tp = positivo verdadeiro fn = falso negativo

F-measure: essa medida, também conhecida por medida F, combina

Precision e Recall. Essa medida é calculada pela seguinte fórmula:

$$F - measure = (2xPrxRc)/(Pr + Rc) \tag{2.1}$$

Onde: Pr = Precision Rc = Recall

Kappa: é uma medida de concordância entre os dados. Esta medida de concordância tem como valor máximo 1, onde este valor 1 representa total concordância e os valores próximos e até abaixo de 0, indicam nenhuma concordância.

Agresti (**??**(**??**), **??** apud **??**(**??**), **??**) sugerem a Tabela **??** para a interpretação dos valores de Kappa.

Essa medida é calculada pela seguinte fórmula:

$$kappa = (po - pe)/(1 - pe) \tag{2.2}$$

po é a probabilidade de concordância

pe é a probabilidade esperada se as proporções forem independentes.

Aplicando-se as medidas de avaliação descritas acima, tem-se:

po = (7/14 + 2/14) = 0.642857

pe = (9/14 * 10/14 + 5/14 * 4/14) = 0.561225

tp = 7

tf = 3

fn = 2

Precision = 7 / (7 + 3) = 0,70

Recall = 7 / (7 + 2) = 0,778

F-measure = 2 * 0,70 * 0,778 / (0,778 + 0,70) = 0,737

Kappa = (0,642857 - 0,561225) / (1 - 0,561225)

Kappa = 0,081632 / 0,438775

Kappa = 0,186

Conforme a Tabela **??**, o valor 0,186 indica baixo grau de concordância.

2.2 Mineração de Texto

Grande parte das informações das empresas encontram-se de forma não estruturada, tais como: documentos digitalizados, e-mails, memorandos, registros de reclamações de clientes, registros de ocorrências, dentre muitos outros. Em busca destas informações para gerar novos conhecimentos há uma técnica que busca descobrir ou minerar conhecimentos neste tipo de informação não estruturada. Esta técnica possui o nome de Mineração de Textos (MT) ou, em inglês, *Text Mining*. Segundo (**??**(**??**), **??** apud **??**(**??**), **??**) a descoberta de conhecimento ocorre por meio de complexas interações realizadas entre homem e uma base de dados, geralmente utilizando uma série heterogênea de ferramentas.

Em (**??**) afirma-se que as três grandes áreas que lidam com informações em grandes bases de dados são: mineração de dados para dados estruturados; extração de informação para dados não estruturados e Recuperação da Informação para textos ou palavras. A tecnologia de mineração de texto serve para identificar os conceitos presentes nos textos. Conceitos representam "entes" do mundo real (entidades, eventos, objetos, sentimentos) e podem permitir entender que temas estão sendo tratados nos textos. Em seguida, a exploração pode utilizar um processo automático de mineração.

Esta mineração pode ser feita analisando-se a distribuição dos conceitos em coleções (a freqüência ou probabilidade com que aparecem) e a relação dos conceitos entre si, para

descobrir associações e dependências.

A MT é uma técnica para a análise de textos, que permite: recuperar informações, extrair dados, resumir documentos, descobrir padrões de associações e regras para classificação. Além disto, é possível realizar análises qualitativas ou quantitativas e desta forma, soluciona grande parte dos problemas relacionados à busca, recuperação e análise de textos. Podemos observar o processo de MT na Figura **??**.

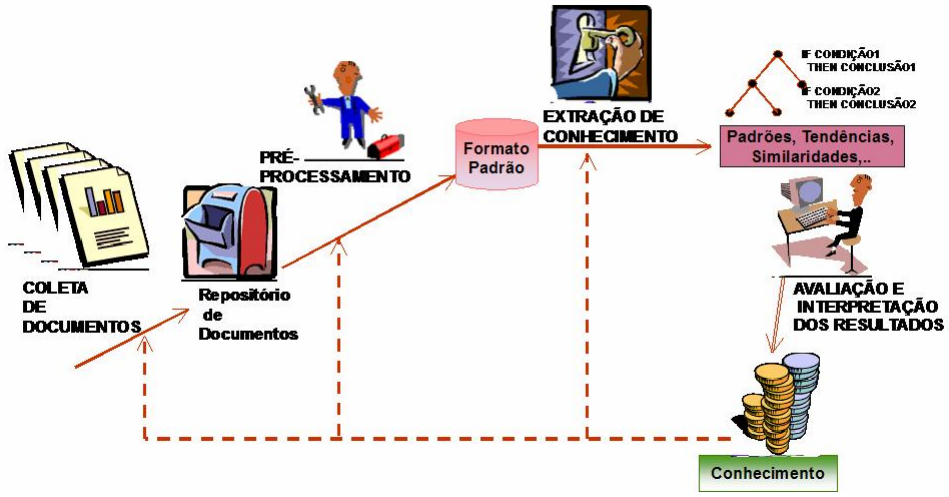

Figura 8: Os passos do processo Mineração de Texto

Os sistemas de informação suporte à MT podem beneficiar os usuários, auxiliando-os a coletar e analisar os dados necessários à tomada de decisão e permitindo com que se posicionem melhor em suas atribuições, aumentando a eficiência e reduzindo erros.

2.2.1 Conceitos básicos sobre Mineração de Texto

Para um melhor entendimento sobre MT é necessário o conhecimento de alguns conceitos básicos para auxiliar no entendimento da técnica de MT. Assim, a presente subseção desse capítulo apresenta esses conceitos e as principais características dessa tecnologia.

Stopwords

Segundo (**??**) *stopwords* "são palavras que ocorrem freqüentemente em textos. Uma vez que elas são muito comuns, sua presença não contribui significativamente para a determinação do conteúdo do documento". Podemos então concluir que elas podem ser descartadas do documento, para fins de MT. As *stopwords* podem ser: os artigos, preposições, pronomes e demais palavras utilizadas para auxiliar na construção sintática das

orações. Para Wives (**??**) a tradução como "palavras negativas", "palavra ferramenta" ou "palavras vazias". Sua remoção contribui ao aumentar a rapidez da operação, pois uma busca que emprega o termo "de", certamente, recupera quase todos os registros em uma base de dados. A Figura **??** sugere uma lista padrão das palavras em português que normalmente se encaixam no conceito de *stopwords*. Caso exista uma palavra que se repita muito na base de dados e que não tenha importância no processo de busca, esta deve ser incluída na lista. Então é possível dizer que a lista das *stopwords* pode variar, dependendo da aplicação/área.

a	desligado	faz	ou
acerca	deve	fazer	outro
agora	devem	fazia	para
algumas	deverá	fez	parte
alguns	direita	fim	pegar
ali	diz	foi	pelo
ambos	dizer	foram	pessoas
antes	dois	horas	pode
ao	dos	iniciar	poderá
apontar	e	início	podia
aquela	é	ir	por
aquelas	ela	irá	porque
aquele	ele	isto	povo
aqueles	eles	ligado	primeiro
aqui	em	maioria	qual
atrás	enquanto	maiorias	qualquer
bem	então	mais	quando
bom	está	mas	quê
cada	estado	mesmo	quem
caminho	estão	meu	quieto
cima	estar	muito	saber
com	estará	muitos	são
como	este	não	sem
comprido	estes	nome	ser
conhecido	esteve	nós	seu
corrente	estive	nosso	somente
das	estivemos	novo	tal
debaixo	estiveram	o	também
dentro	eu	onde	tanto
desde	fará	os	tem

Figura 9: Lista de *Stopwords*

Corpus

Corpus, segundo Sardinha (**??**), "é um conjunto de dados lingüísticos, organizados seguindo alguns critérios, de maneira que sejam representativos do conjunto de dados, armazenados de tal modo que possam ser processados por computador, com a finalidade de gerar resultados úteis para a descrição e análise". Existem dois tipos de *corpus*:

- um *corpus* de estudo, representado em uma lista de freqüência de palavras. O corpus de estudo é aquele que se pretende descrever;
- um *corpus* de referência, também formatado como uma lista de freqüência de palavras. Também é conhecido como "*corpus* de controle", e funciona como termo de comparação para a análise. A sua função é a de fornecer uma norma com a qual se fará a comparação das freqüências do *corpus* de estudo. A comparação é feita através de uma prova estatística selecionada pelo usuário. As palavras cujas freqüências no *corpus* de estudo forem significativamente maiores segundo o resultado da prova estatística são consideradas chaves, e passam a compor uma listagem específica de palavras-chaves.

O *corpus* de referência não deve conter documentos relacionados ao tema a ser analisado, pois ao comparar-se o corpus de referência com o corpus de estudo, a diferença entre eles será uma lista de palavras-chaves ou keywords.

Keywords

Keywords ou palavras-chave são palavras cuja freqüência é estatisticamente diferente no *corpus* de estudo em relação ao corpus de referência. Segundo (**??**), "para se analisar quais são as keywords necessita-se de dois elementos básicos: corpus de estudo e corpus de referência". Estas palavras passam a compor uma listagem específica de palavras-chave, mais significativas no texto.

A ferramenta proposta se utiliza de parte deste conceito para identificar e classificar as palavras mais significativas para o documento. Desta forma é possível afirmar que o estudo da freqüência das palavras pode gerar uma lista das palavras.

Em (**??**) é sugerido o seguinte algoritmo para extração de palavras-chave:

1. Selecionar o primeiro item na lista de palavras do corpus de estudo.
2. Procurar por este item na lista de palavras do corpus de referência.
3. Se o item constar no *corpus* de referência, ir para o passo a seguir, senão passe a ir para o passo sete.
4. Comparar as freqüências através de uma prova estatística escolhida pelo usuário.
5. Se o resultado da comparação for estatisticamente significativo, copiar esta palavra para uma nova lista, e chamá-la de lista de palavras-chave.

6. Repetir este procedimento até o último item da lista de palavras do *corpus* de estudo.

7. Se um item constante da lista de palavras do *corpus* de estudo não aparecer na lista de palavras do *corpus* de referência, assumir freqüência zero para este item no *corpus* de referência.

8. Executar os passos quatro, cinco, e seis.

Collocations

Segundo (**??**) "as colocações ou expressões compostas são agrupamentos de palavras onde o significado do todo é a soma dos significados das partes mais algum componente semântico adicional não previsto pelas partes".

Manning & Schütze (**??**) definem *collocations* como "uma expressão que consiste em duas ou mais palavras que correspondem a algum modo convencional de dizer alguma coisa".

Choueka (**??**) afirma:

> "uma colocação é definida como uma seqüência de duas ou mais palavras consecutivas que têm características de uma unidade sintática e semântica, e cujo significado ou conotação exato e não-ambíguo não possa ser derivado diretamente a partir do significado ou conotação de seus componentes".

Para (**??**(**??**), **??** apud **??**(**??**), **??**) "colocações de uma dada palavra são afirmações dos lugares comuns ou habituais daquela palavra."

Um exemplo disso está na expressão "chutar o balde". Apesar de não está falando sobre baldes, mas é uma forma de expressão utilizada no idioma português para expressar que não se está preocupado com o que vai acontecer depois de determinada atitude, demonstrando que a semântica da frase está sendo levado em conta.

Stemming

Porter (**??**) afirma que "*stemming* consiste em converter cada palavra para seu 'radical' ('*stem*'), isto é, uma forma neutra com respeito a *tag-of-speech* e inflexões verbais plurais. Por exemplo, as palavras '*learning*' e '*learned*' são ambas convertidas para o *stem* '*learn*'. Segundo Chaves (**??**) "*stemming* consiste em reduzir todas as palavras ao mesmo *stem*, por meio da retirada dos afixos da palavra, permanecendo apenas a raiz dela".

O propósito, segundo (**??**) é:

> "Chegar a um *stem* que captura uma palavra com generalidade suficiente para permitir um sucesso na combinação de caracteres, mas sem perder muito detalhe e precisão. Um exemplo típico de um *stem* é 'conect' que é o *stem* de 'conectar', 'conectado' e 'conectando'. Dois erros típicos que costumam ocorrer durante o processo de *stemming* são *overstemming* e *understemming*. *Overstemming* se dá quando a cadeia de caracteres removida não é um sufixo, mas parte do *stem*. Por exemplo, a palavra gramática, após ser processada por um *stemmer*, é transformada no *stem* grama. Neste caso, a cadeia de caracteres removida eliminou parte do *stem* correto, a saber 'gramát'. Já *understemming* ocorre quando um sufixo não é removido completamente. Por exemplo, quando a palavra 'referência' é transformada no *stem* 'referênc', ao invés do *stem* considerado correto 'refer'.

Em 1980, Martin Porter (**??**) desenvolveu um algoritmo que tem por objetivo o tratamento de *stemming* para a língua inglesa. Tem sido adaptado para várias línguas latinas, tais como espanhol e português.

Limpeza dos dados (Data Cleaning)

Para construir bons modelos precisa-se de dados "limpos". Entretanto, os dados em muitas organizações possuem baixa qualidade. Valores ausentes, valores ilegais, combinações inexistentes e erros de grafia podem alterar seus resultados. Os recursos de transformações e limpeza nos dados aumentam o valor desses dados.

Em (**??**) afirma-se que "a qualidade dos dados é essencial para a obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração". As atividades de obtenção e limpeza dos dados normalmente consomem mais da metade do tempo dedicado ao projeto. Porém a limpeza dos dados pode evitar que a consolidação dos dados sejam distorcidos. Geralmente são erros pequenos e simples, onde uma letra é adicionada, trocada ou omitida. São erros difíceis de serem encontrados em um conjunto de dados pela pequena diferença na ortografia. Segundo Han & Kamber (**??**) as tarefas para limpeza da base são:

1. preencha os valores que estiverem faltando;

2. identifique outliers, ou seja, valores muito distantes da média e retire os ruídos dos dados;

3. verifique a consistência dos dados;

4. resolva a redundância causada pela integração dos dados.

2.2.2 Descoberta reativa e pro-ativa

Existem dois modos de descoberta: reativa ou pro-ativa. Na descoberta reativa é necessário que se entenda qual o interesse ou objetivo do usuário para limitar o espaço de busca na entrada ou filtrar os resultados na saída (**??**). O usuário tem uma idéia vaga do que pode ser a solução ou de onde pode encontrá-la. O usuário possui algumas hipóteses iniciais que serão utilizadas para direcionar o processo de descoberta. Neste caso, é necessário que haja algum tipo de pré-processamento, para selecionar atributos ou valores de atributos.

Na descoberta pro-ativa, ao contrário da reativa, a solução do problema é encontrada automaticamente, sem a intervenção do usuário. Segundo Loh & Oliveira (**??**), uma expressão comum para definir o modo pro ativo é: "diga-me o que há de relevante nesse conjunto de dados".

Nesta dissertação é utilizado o modo reativo de descoberta.

2.2.3 Tarefas mais comuns de descoberta de conhecimento em textos

Encontra-se na literatura as seguintes tarefas de descoberta de conhecimento de textos: classificação e categorização, sumarização, *clustering*, regras de associação e recuperação de informação (RI) ou *Information Extraction* (IE). Qualquer um dos métodos de descoberta de MD pode ser aplicado nos textos. Nesta dissertação optou-se pelo método de recuperação de informações para auxiliar no processo de mineração dos textos.

O método de RI auxilia pessoas a encontrar informações relevantes em documentos não estruturados. Este método é particularmente interessante quando da utilização do modo de descoberta reativa, pois neste caso, como já foi dito, é necessário que se conheça o interesse ou objetivo do usuário. Sabendo-se qual é o interesse do usuário é possível montar um domínio do problema permitindo diminuir a possibilidade de haver um problema chamado de "sobrecarga de informações" (*information overload*) que acontece quando

o usuário recebe um conjunto muito grande de documentos que não satisfazem a sua pergunta. Um exemplo típico quando utilizamos este método sem a utilização do domínio é a utilização de sistemas de busca (*search engines*) na Web, quando uma pesquisa pode retornar uma quantidade imensa de respostas possíveis, obrigando o usuário a fazer uma análise de cada resposta encontrada.

Na seqüência é feita uma descrição do método de RI para descoberta de conhecimento em textos:

2.2.4 Recuperação de Informação

Recuperação de informação estuda o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos. Um Sistema de Recuperação de Informação (SRI), segundo (**??**) pode ser estruturado conforme a Figura **??**.

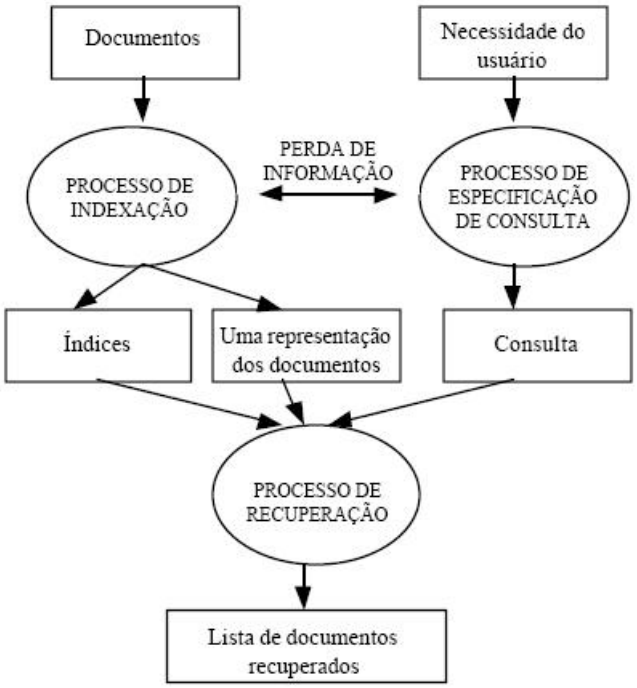

Figura 10: Componentes de um sistema de recuperação de informação

2.2.5 Componentes de um Sistema de Recuperação de Informação

Os componentes de um SRI incluem documentos, necessidades do usuário que monta uma consulta (pergunta) e o processo de recuperação que, baseado nas estruturas de dados e da pergunta formulada, recupera uma lista de documentos considerados relevantes. O

processo de indexação envolve a criação de estruturas de dados associados à parte textual dos documentos. Estas estruturas podem conter dados sobre características dos termos na coleção de documentos, tais como a freqüência de cada termo em um documento, como discutidas em (**??**). O processo de especificação da consulta ou pergunta geralmente pode ser uma tarefa difícil. Há com freqüência uma diferença semântica entre a sua necessidade e o que é expresso em sua pergunta. Essa diferença é gerada pela limitação do conhecimento do usuário sobre o domínio da pesquisa e pelo formalismo existente na linguagem de consulta. O processo de recuperação consiste na geração de uma relação de documentos recuperados para responder a pergunta formulada pelo usuário. Os índices são construídos para um conjunto de documentos e são utilizados para acelerar esta tarefa. Além disso, a relação de documentos recuperados é classificada em ordem decrescente de um grau de relevância entre o documento e a pergunta.

2.2.6 Avaliação de sistemas de recuperação de informação

Os SRI podem ser avaliados através de consultas que fazem parte de uma conjunto referência. Para este conjunto referência é fornecido um conjunto ideal de documentos resposta. Este conjunto ideal de documentos resposta é criado por especialistas nos temas envolvidos. Assim é possível avaliar o SRI através da comparação das respostas geradas por este sistema e o conjunto ideal de respostas. Feito isto, o conjunto resultado é examinado e comparado com o conjunto ideal, obtendo-se dois índices de avaliação: precisão (*precision*) e revocação (*recall*).

A precisão avalia se os documentos recuperados são todos relevantes e a revocação avalia se todos os documentos relevantes são recuperados.

Sejam:

- N é conjunto de documentos relevantes identificados por especialistas
- I o conjunto de documentos recuperados pelo sistema
- P é a precisão do sistema
- R é a revocação do sistema

A precisão e dada por :

$$P = \frac{|N \bigcap I|}{|I|} \tag{2.3}$$

A revocação é dada por:

$$R = \frac{|N \bigcap I|}{|N|} \tag{2.4}$$

Um exemplo:

- documentos relevantes: 2, 33, 55, 23, 99.
- um sistema recupera o conjunto resultado: {55, 25, 99, 6, 2, 8, 21, 7, 29, 14, 33, 40, 36, 22, 81, 23 }.

O nível de revocação 20% é atingido quando encontramos o primeiro documento relevante(55), a precisão é de 1/1 = 100%. Para revocação de 40% a precisão é igual a 2/3 = 66%.

2.2.7 Modelos Clássicos

Os modelos clássicos utilizados no processo de RI (booleano, vetorial e probabilístico) apresentam estratégias de busca de documentos relevantes para uma consulta. Estes modelos consideram que cada documento é descrito por um conjunto de palavras chaves, chamadas termos de indexação. Associa-se a cada termo de indexação t_i em um documento d_j um peso $w_{ij} \geq 0$, que quantifica a correlação entre os termos e o documento. Além dos modelos clássicos, modelos muito mais avançados de recuperação de informação tem sido propostos ao longo dos anos, dentre estes, destacam-se modelos baseados em bases de conhecimento, lógica fuzzi e redes neurais.

Modelo Booleano

O modelo booleano é um dos modelos clássicos que considera uma consulta como uma expressão booleana convencional, que liga seus termos através de conectivos lógicos AND, OR e NOT. No modelo booleano um documento é considerado relevante ou não relevante a uma consulta, não existe resultado parcial e não há informação que permita a ordenação do resultado da consulta. Este modelo é muito mais utilizado para recuperação de dados do que para recuperação de informação. É bom para quem entende bem de álgebra booleana, mas o usuário, na maioria dos casos, não entende.

As consultas são construídas como uma combinação dos conjuntos que descrevem todas as possibilidades para o conjunto resposta da consulta, chamadas de min-termos. Geralmente, para n termos, temos: k = 2 ^{n} min-termos e 2 ^{k} consultas. Por exemplo, para

3 termos são 8 min-termos e 2^8 (256) possíveis consultas. O tamanho da base de dados afeta tanto as estratégias de consultas, quanto os resultados obtidos usando o método booleano. Este modelo é altamente utilizado em sistemas comerciais.

Vantagens do modelo booleano:

- expressividade completa se o usuário souber exatamente o que quer;
- é facilmente programável e exato.

Desvantagens do modelo booleano:

- pessoas lidam com conhecimento parcial;
- Saída pode ser nula, ou haver overload;
- Saída não é ordenada.

Algumas formas de tentar melhorar os resultados gerados nesse modelo:

- atribuindo pesos aos termos. A carga semântica dos termos é completamente diferente, quando se tem, por exemplo, uma consulta com dois termos ou duas consultas distintas com um termo cada;
- usando conjuntos fuzzy. A pertinência ou não de um elemento a um conjunto varia entre 0 e 1, não é exata;
- Categorização em RI. Dividir a consulta em classes e conceitos, tentar encontrar os documentos baseados nos conceitos.
- Passage retrieval. O conjunto de termos a ser procurado deve aparecer o mais próximo possível, por exemplo, da mesma página (uma possível passagem). É uma técnica mais eficiente que a de RI, porém muito mais difícil de ser implementada. A proximidade é importante. Ordenando a saída. Uma vez que haja alguma forma de determinar que termos são mais importantes para determinada consulta é possível ordenar o resultado.

Modelo Vetorial

O modelo espaço-vetorial (ou simplesmente vetorial) foi desenvolvido por Gerard Salton (**??**), para ser utilizado num SRI chamado SMART. No modelo vetorial, cada documento é representado como um vetor de termos e cada termo possui um valor associado

que indica o grau de importância (peso - weight) deste no documento. Em outras palavras, cada documento possui um vetor associado que é constituído por pares de elementos na forma $\{palavra_1, peso_1, palavra_2, peso_2, ..., palavra_n, peso_n\}$.

Segundo (**??**) o peso de um termo em um documento pode ser calculado de diversas formas. Os pesos são usados para computar a similaridade entre cada documento armazenado e uma consulta feita pelo usuário. Esses métodos de cálculo de peso geralmente se baseiam no número de ocorrências do termo no documento (freqüência).

Vantagens do modelo vetorial:

- atribuir pesos aos termos melhora o desempenho;
- é uma estratégia de encontro parcial (função de similaridade), que é melhor que a exatidão do modelo booleano;
- os documentos são ordenados de acordo com seu grau de similaridade com a consulta.

Desvantagens do modelo vetorial:

- ausência de ortogonalidade entre os termos. Isso poderia encontrar relações entre termos que aparentemente não têm nada em comum;
- é um modelo generalizado;
- um documento relevante pode não conter termos da consulta.

Modelo Probabilístico

O modelo probabilístico possui esta denominação justamente por trabalhar com conceitos provenientes da área de probabilidade e estatística. Neste modelo, os termos indexados dos documentos e das consultas não possuem pesos pré-definidos. A ordenação dos documentos é calculada pesando dinamicamente os termos da consulta relativamente aos documentos. É baseado no princípio da ordenação probabilística (*Probability Ranking Principle*). Nesse modelo, busca-se saber a probabilidade de um documento D ser ou não relevante para uma consulta C_a. Tal informação pode ser obtida assumindo-se que a distribuição de termos na coleção seja capaz de informar a relevância provável para um documento qualquer da coleção.

2.2.8 Relação das principais ferramentas de Mineração de Texto

Segundo (**??**) as principais ferramentas de mineração de texto são:

TextAnalyst

Fornecedor: *Megaputer Intelligence, Inc.*

É a principal ferramenta que disponibiliza sumarização de texto. É um excelente programa, fácil de utilizar que trabalha com texto não-estruturado, como artigos e informes e produz um sumário preciso. Esta tecnologia de sumarização independe da linguagem e está disponível em diversas línguas, como inglês, francês, alemão, espanhol, italiano, russo e holandês. Embora *TextAnalyst* seja um pequeno componente em um grande sistema de software de inteligência competitiva, possibilita um ganho inestimável de tempo para o time de IC que precisa coletar e ler uma grande quantidade de documentos escritos diariamente. Pode também ajudar na etapa de relatórios, pois fornece sumários para os executivos, que decidirão se querem ou não ler a análise completa. Além de sumarização, esta ferramenta realiza agrupamentos e possui um sistema de pergunta-resposta.

Market Signal Analyzer

Fornecedor: *Docere Intelligence, Inc.*

Potente e flexível ferramenta para Inteligência Competitiva, que suporta todo o ciclo: planejamento, coleta, estruturação, análise qualitativa da informação e geração de relatórios. Estrutura baseada em matriz para coletar e organizar informação qualitativa a fim de identificar e relatar tendências e/ou eventos que podem impactar empresas. Este estilo funciona como sistema de avisos antecipados. A coleção e análise da informação são, na maior parte, manual. Pouca coisa é automatizada ou dinâmica. Entretanto, a estrutura é ideal para que uma equipe razoavelmente nova do time de Inteligência Competitiva e que esteja interessada em automatizar todo o processo.

C-4-U Scout

Fornecedor: *C-4-U Ltd*

Ferramenta que utiliza *queries* que podem extrair dinamicamente informações alvo na internet, podendo também buscar em outras fontes de dados não estruturados. Suporta uma coleção de informação textual qualitativa, bem como dados quantitativos. Existe em

três diferentes configurações, a fim de satisfazer diferentes necessidades de negócios e financeiras. Uma vez extraídas as informações, elas podem automaticamente ser novamente formatadas em banco de dados, planilhas, XML, etc.

Knowledge Works

Fornecedor: *Cipher*

Desenvolvido especificamente para Inteligência Competitiva (IC) nas indústrias, fornecendo: Fluxo de trabalho centrado ao redor de "*Key Intelligence Topics*" e "*Key Intelligence Questions*". Coleção automatizada de informações publicadas a partir de fontes de dados internas e externas. Permite a entrada de informações primárias no sistema após a análise.

Clear Research Suite

Fornecedor: *ClearForest Corporation*

Uma das melhores ferramentas revisadas que fazem aplicações de análise, extração de características, visualização de inter-relações complexas entre empresas, pessoas, eventos etc. no mundo dos negócios. É considerada avançada nas fases nas quais esta ferramenta se propõe a automatizar. O motor de extração de informação pode, dinamicamente, identificar relacionamentos entre pessoas, companhias e grandes repositórios de textos não estruturados, incluindo novos fontes, páginas da Web e informes internos. O monitoramento é baseado em Web e notificação em tempo real de eventoschave em negócios. Os múltiplos produtos do *ClearResearch Suíte* (*ClearReserach*, *ClearTags*, *ClearSight*, *ClearEvents*, *ClearCharts*) fornecem uma visão única destes relacionamentos extraídos.

BrandPulse

Fornecedor: *Intelliseek Inc.Planetfeedback*

Permite buscar opiniões e tendências na internet, monitorando bases de dados públicas, quadros de discussão, opiniões, boatos e oportunidades em tempo real. Identifica e reage rapidamente às mudanças de necessidades do consumidor e suas opiniões. Monitora a percepção da companhia e produtos 24horas por dias, identificando suas forças, fraquezas e desempenho. Reduz o monitoramento manual e relata os custos. Informa antecipadamente rumores e problemas antes de ser afetado pelo estágio de crise. Além de capturar

novas idéias, mensagens específicas, identificando os usuários, a fim de melhorar o esforço de *marketing*. *BrandPulse* objetiva ser um administrador da marca, desenvolvedor de produtos e relações públicas profissional, que pode beneficiar a partir de um profundo entendimento da percepção do consumidor, satisfação, comportamento das palavras, fatores competitivos bem como tendências da indústria.

TrackEngine

Fornecedor: *NexLabs Pte Ltd.*

Eficiente programa que traça rotas em páginas da Web, podendo monitorar *Web sites* corporativos, salas de bate-papo e quadros incorporados em mensagens. Pode alertar o usuário proativamente de qualquer novo conteúdo, através de um *e-mail* alerta. É o núcleo de um pacote de IC bem maior, chamado *IntelligenceWork*. Este pacote fornece mais funcionalidades e estruturas para rotear e recolher informação. Monitora e alerta potencialidades de corporações através de mensagens 5nteligentes, informando quando o conteúdo de um *site* é atualizado, a partir de uma lista de *sites* previamente indicados como mais importantes.

Strategy

Fornecedor: *Strategy Software, Inc.*

A força de *STRATEGY* encontra-se em seu suporte à informação organizada a partir de muitas fontes, principalmente táticas, para criar uma base multidimensional para a análise eficaz e a tomada de decisão. O produto aparece particularmente bem adaptado para os profissionais do time de Inteligência Competitiva que apóiam vendas e *marketing* de clientes dentro da firma. *STRATEGY* fornece ao usuário um meio de organizar informações diferentes de maneira estruturada. O software tem recursos para comparar uma grande variedade de matrizes e outros tipos de relatórios de *benchmarking*, que podem ser disseminados por uma variedade igualmente variada de canais. Auxilia grupos de usuários na criação e manutenção de uma consciência coletiva de IC, onde cada um dos usuários pode contribuir e aprender a partir dela. Permite que multiusuários obtenham todos os tipos de informações sobre a sua companhia, competidores, indústria e a economia num caminho lógico e estruturado.

PlanBee

Fornecedor: *Thoughtshare Communication Inc.*

Permite consolidar *Web pages*, documentos texto, arquivos de imagem, PDFs, arquivos de áudio e arquivos de vídeo. Ajuda aos usuários a "empacotar" a informação baseada na Web, facilitando desse modo sua disseminação. O time de IC pode relacionar *Web pages*, comentários e anexados em um único arquivo, chamada um buzPak, que pode ser enviado por *e-mail*. *PlanBee* permite criar anotações no *buzPak*, que servirão como *tour* para o usuário. Deste modo, o time de IC pode ter algum controle sobre a inteligência revista.

Wincite

Fornecedor: *Wincite Systems LLC*

Organiza informações armazenadas em um BD relacional. Análise estratégica e planejamento, gerenciamento do produto, pesquisa de mercado. Captura organiza, distribui inteligência na empresa. Abrangente ferramenta de banco de dados de inteligência competitiva que pode aumentar muitos passos no ciclo de inteligência. *E-Wincite* permite acessar remotamente os dados na base de dados do *Wincite* através de um *browser* de internet. A força de *Wincite* encontra-se em suas estruturas analíticas e nas características do relatório, ambos podem ser valiosos para os analistas de Inteligência Competitiva.

Wisdom Builder

Fornecedor: *Wisdom Builder, LLC*

Auxilia a encontrar e extrair em grande quantidade de informação. A força do *Wisdom Builder's* encontra-se em sua arquitetura colaborativa integrada que pode dinamicamente ser "costurada" pelo usuário, permitindo que o usuário tenha uma grande flexibilidade durante o processo de pesquisa. O *Wisdom Builder's* tem um foco único em encontrar relacionamentos ocultos entre eventos, pessoas, lugares, produtos e organizações em texto não-estruturado (isto é, artigos de notícia, *press release*, etc.). A maioria das demais ferramentas realiza a comparação com outros produtos em uma base de dados estruturada, pré-processada. O *Wisdom Builder's* possui uma funcionalidade formidável de busca. Os resultados podem ser analisados e relatados por uma grande variedade de formatos.

S-Miner

Fornecedor: Coordenação dos Programas de Pós-graduação de Engenharia, COPPE, UFRJ

O S-Miner (**??**) foi desenvolvido para minerar textos. Trabalha com português do Brasil. Inicialmente, o texto é submetido ao algoritmo de geração de *tokens*. A tokenização consiste na identificação de palavras (ou *tokens*). Esta técnica sugere que os *tokens* sejam definidos como uma *string* de caracteres alfanuméricos sem espaços. Após a quebra do texto em *tokens*, o processo prossegue com a retirada das palavras que não possuem relevância significativa no texto, as chamadas *stopwords*. O conjunto de *stopwords* que serão retirados do texto compõe a *Stop List*. Esta lista de palavras irrelevantes é fortemente dependente da língua e do contexto utilizados.O S-Miner suporta inglês e português (brasileiro).

Eureka

Fornecedor: Leandro Krug Wives

A ferramenta Eureka (**??**) se propõe a analisar um conjunto de textos não formatados e a identificar e agrupar aqueles considerados semelhantes semanticamente. Dentre as facilidades oferecidas pela ferramenta estão a possibilidade de configuração de listas de *stopwords* e da escolha de um entre quatro algoritmos de agrupamento.

3 Ambientes Virtuais de Aprendizagem

Em (3) é definido *ambientes de aprendizagem* como sistemas de ensino e aprendizagem integrados e abrangentes capazes de promover o engajamento do aluno.

Neste processo de ensino e aprendizagem, os estudantes devem ser sujeitos do processo de aprendizagem. Para isso, devem ser criadas situações de ensino e aprendizagem nas quais eles mesmos possam organizar seus estudos. Os próprios de estudos devem ser iniciados por meio de discussão e interação que é chamado de princípio do estudo por meio de comunicação e interação em (4). Segundo Landim (5) a interatividade envolve as mediações que constituem o tratamento dos conteúdos e das formas de expressão e relação comunicativa, que possibilitam a aprendizagem à distância.

As novas tecnologias, principalmente o computador e a Internet, têm proporcionado a criação de comunidades virtuais, cujos membros podem comunicar-se síncrona ou assíncronamente e sem estarem necessariamente no mesmo lugar. Esses membros podem interagir das mais diversas formas e com os mais variados objetivos.

Um Ambiente Virtual de Aprendizagem (AVA) tem como objetivo apoiar comunidades, organizadas das mais variadas formas, visando o desenvolvimento de atividades individuais e/ou coletivas, tais como o esclarecimento de dúvidas, desenvolvimento de trabalhos em grupo, entre outras. A interatividade em AVA é fundamental para que os alunos possam organizar suas idéias, compartilhar seus conhecimentos tornado-se sujeitos autônomos de sua aprendizagem (5).

Um grande volume de dados é gerado por meio do uso de diferentes ferramentas de interação. No entanto, é comum não achar nos diferentes AVA tratamento e estruturação adequados para esses dados. Este fato faz com que informações que possam existir não sejam aproveitadas (6), perdendo assim uma fonte de informações que poderia estar agregando conhecimento ao AVA. Isso sugere com que o AVA possa ser um ambiente propício

para a mineração de dados, uma vez que essa técnica tem como objetivo extrair conhecimento implícito a partir de um grande volume de dados.

Disponibilizar um ambiente virtual de aprendizagem que enriqueça a cooperação e a interatividade é uma das metas atuais à qual a literatura mostra que estão dirigindo esforços.

3.1 Ferramentas de comunicação em Ambientes Virtuais de Aprendizagem

As tecnologias de informação e comunicação estão se tornando ferramentas que cada vez interativas e distribuídas, quando empregadas em AVA, proporcionam aos alunos e professores um conjunto de meios para que possam compartilhar de informações e recursos. As potencialidades da Internet e dos serviços suportados por esta, estão sendo utilizadas no processo de ensino e aprendizagem, não só pela influência da elevada quantidade e variedade de meios que disponibilizam, mas também, pelas múltiplas perspectivas de abordagem que proporcionam.

Os ambientes de aprendizagem baseados na Web surgem como um ambiente flexível que favorece o modo de trabalhar de cada aluno, de forma nomeada, de poder trabalhar ao seu próprio ritmo em qualquer lugar e a qualquer hora, de atuar de modo individual ou em grupo, de aprender a relacionar-se com os outros, comunicar, colaborar e partilhar com outros da sua comunidade de aprendizagem.

Nos ambientes AVA são utilizados vários tipos de ferramentas que possibilitam esta forma de utilização pessoal. Estas ferramentas possuem características que a tornam específicas para cada etapa no processo ensino-aprendizagem. Segundo Perrone (7), estas ferramentas podem ser classificadas por aplicabilidade da seguinte forma:

- Apresentação do Ambiente;
- Socializantes;
- Comunicação do curso;
- Acesso ao conteúdo;
- Construção coletiva;
- Avaliação dos participantes;

- Gerência do ambiente;

Nas ferramentas de comunicação do curso estão inseridas as ferramentas como *chat* ou bate-papo, forum, listas de discussão e conferências. A ferramenta *chat* é utiliza a comunicação síncrona e pode envolver muitas pessoas, possibilitando um bom nível de interatividade. Este trabalho propõem uma ferramenta para melhorar este nível de interatividade, possibilitando, no momento desta comunicação, uma consulta a uma base de conhecimento que pode facilitar o contato entre seus participantes.

3.2 Ambiente Virtual de Aprendizagem Teleduc

O TelEduc [1] (8) é um ambiente para a criação, participação e administração de cursos na Web. Ele foi concebido no ano de 1998 tendo como alvo o processo de formação de professores para informática educativa, baseado na metodologia de formação contextualizada desenvolvida por pesquisadores do NIED (Núcleo de Informática Aplicada à Educação) da Unicamp.

O TelEduc foi desenvolvido de forma participativa, ou seja, todas as suas ferramentas foram idealizadas, projetadas e depuradas segundo necessidades relatadas por seus usuários. Com isso, ele apresenta características que o diferenciam dos demais ambientes para educação a distância disponíveis no mercado, como a facilidade de uso por pessoas não especialistas em computação, a flexibilidade quanto a como usá-lo, e um conjunto enxuto de funcionalidades.

O mesmo foi concebido tendo como elemento central uma ferramenta que disponibiliza atividades. Isso possibilita a ação onde o aprendizado de conceitos em qualquer domínio do conhecimento é feito a partir da resolução de problemas, com o subsídio de diferentes materiais didáticos como textos, *software*, referências na Internet, dentre outros, que podem ser colocadas para o aluno usando ferramentas como: material de apoio, leituras, perguntas freqüentes, etc.

A intensa comunicação entre os participantes do curso e a ampla visibilidade dos trabalhos desenvolvidos também são pontos importantes, por isso foi desenvolvido um amplo conjunto de ferramentas de comunicação como o correio eletrônico, grupos de discussão, mural, portfólio, diário de bordo, bate-papo.

[1] Teleduc - Núcleo de Informática Aplicada à Educação - Unicamp - http://www.nied.unicamp.br/oea/soft/teleduc.html-2006

O TELEDUC hoje é utilizado por várias instituições de ensino: universidades federais, particulares, centros tecnológicos de educação (CEFET), faculdades e associações.

3.3 Trabalhos Relacionados

Com o uso do computador, tornou-se possível capturar algumas características do aprendiz, à distância, e analisá-las de uma maneira análoga ao comportamento de um aluno de um curso presencial. A linguagem corporal, o grau de interesse, a participação, o comportamento social, podem ser vistos pela ótica computacional, considerando, basicamente, as iterações do aluno com o ambiente de ensino. A freqüência de sua participação em listas de discussão, conferências, fórum, *e-mails* e *chats*, por exemplo, podem retratar sua sociabilidade.

A seguir é apresentado uma revisão de trabalhos que tratam o acompanhamento dos alunos em AVA, tentando transcrever características da educação presencial para o EaD.

Em (9), Lopes propõe-se uma estratégia para o acompanhamento do aprendizado na EaD baseada nas práticas de acompanhamento do ensino presencial, acrescida da tática de análise de dados, onde fatores do acompanhamento podem ser relacionados para se verificar a aprendizagem de forma mais elaborada através da geração de um novo conhecimento descoberto com a utilização de ferramentas de mineração de dados. Apresenta-se o MIDAS-POETA, um sistema de apoio a decisão baseado em mineração de dados para o sistema Portfolio-Tutor.

Em (10), Machado e Becker apresenta-se um estudo sobre a utilização de técnicas da mineração de dados, aplicadas sobre dados Web. O trabalho descreve um estudo de caso sobre a mineração do uso da WEB no contexto da educação a distância, no qual, auxilia na identificação de padrões de uso dos dados WEB através das interações dos alunos em um ambiente virtual de aprendizagem. Este estudo de caso envolve um curso baseado na WEB disponibilizado pela PUCRS Virtual.

O artigo de Oliveira e Garcia (11) descreve testes realizados e os resultados obtidos com a aplicação das etapas iniciais do processo de descoberta de conhecimento e a técnica de regras de associação da etapa de mineração de dados. Elas foram utilizadas sobre os dados relacionados ao Questionário Sócio-Econômico-Cultural aplicado durante o Processo Seletivo do Centro Universitário de Formiga no ano de 2004. O objetivo maior do projeto relatado é encontrar informações úteis que possam se transformar em conhecimento estratégico e que possam estar escondidas dentro da base de dados do Processo

Seletivo da Universidade de Fortaleza (UNIFOR).

Em (1), Silva propõe-se um método para o acompanhamento do aprendizado em cursos a distância. Este método está baseado na identificação, estruturação e mineração das informações relevantes das interações do estudante com ambiente de aprendizado e do seu desempenho no curso. O objetivo do trabalho é prover meios para que se possa obter padrões que caracterizem o comportamento dos alunos, os quais servirão como orientações para tornar mais eficiente o acompanhamento do aluno.

No trabalho de Oeiras (12) é proposto o ACEL, um ambiente para o ensino / aprendizagem de línguas a distância. Este ambiente utiliza a Internet como tecnologia para mediar o contato entre professores e alunos, sendo composto de dois sub-ambientes integrados. O primeiro, denominado ambiente do professor, tem definidas as ferramentas computacionais para um professor preparar e operar cursos de línguas para a rede. Esses cursos a distância são acessados pelos alunos através do ambiente de curso que possui recursos de apoio, de comunicação e o material didático. O protótipo implementado foi testado através de um curso de leitura e produção escrita de Português para Estrangeiros cujo público-alvo foi alunos falantes de Espanhol.

A tecnologia de agentes, desenvolvida para aplicações de Inteligência Artificial, tem se mostrado uma das soluções mais adequadas para implementação do sistema de acompanhamento do aluno em ambientes telemáticos. Estão descritas a seguir algumas propostas para o uso dessa técnica.

Em (13), Jaques e Oliveira propõem uma arquitetura Multi-Agente para monitorar os principais mecanismos de comunicação em um ambiente telemático de ensino, entre os quais estão lista de discussão, *newsgroups* e *chat*. A tarefa dessa agência ou sociedade, isto é, uma coleção de agentes trabalhando em conjunto, seria de coletar dados a partir das discussões que se encontram em andamento, analisar quantitativamente esses dados e transmitir tais informações ao professor. Essa sociedade possui quatro agentes: um agente para coletar informações em cada mecanismo de comunicação (*newsgroups*, *chat* e lista) e um agente do professor para reunir as análises dos demais agentes. As análises dos agentes coletores estão baseadas na identificação de três tipos de associações: aluno-aluno, aluno-assunto, aluno-aluno-assunto.

Já em (14), Menezes, Fuks e Garcia propõem uma agência com três agentes assistentes de tarefa para suportar a avaliação informal (acompanhamento) no ambiente AulaNet da Puc-Rio. As interações dos alunos com o ambiente são monitoradas por um agente que cria um histórico da navegação do aluno e percorre as listas de discussão, a fim de

verificar a participação de alunos. Outro agente auxilia o professor na consulta ao relação histórica de interações, ao modelo do aluno e a uma base de conhecimentos responsável pela interpretação desse log de interação. Esse agente também é capaz de confrontar as informações decorrentes dos processos de avaliação informal com as informações resultantes dos processos de avaliação formal do AulaNet. O terceiro agente é capaz de indicar possíveis distorções no design instrucional, refletidas em decorrência do comportamento verificado nos aprendizes.

A ferramenta descrita por Guedes, Viccari e Damico em (15), está relacionada ao Ambiente Multiagente de Ensino-Aprendizagem(AME-A), no qual os agentes que o compõem preocupam-se em ensinar e/ou aprender. Esta ferramenta tem por objetivo possibilitar que diversos aprendizes e professores se comuniquem, através da Internet e discutam assuntos determinados por um professor. Procurando auxiliar a tarefa do professor em determinar se os aprendizes estão realmente adquirindo conhecimento, desenvolveu-se uma ferramenta para analisar as interações dos aprendizes. O algoritmo desenvolvido utiliza um dicionário de palavras/frases-chaves relacionadas ao assunto em questão, referentes a tópicos que deveriam ser discutidos e/ou fazer parte das conclusões dos alunos. Ao ser ativado, o software identifica os aprendizes e suas respectivas interações e as armazena em uma base de dados; em seguida, avalia as interações de cada aprendiz, verificando a freqüência com que este utiliza as palavras-chave. O software permite também a classificação de todas as palavras/frases empregadas durante a reunião.

Já em (16), Jaques e Viccari propõem um agente inteligente para fornecer suporte emocional ao aluno, motivando-o, fazendo-o acreditar em suas próprias habilidades e promovendo um estado de espírito mais positivo no aluno que, de acordo com psicólogos e pedagogos, é melhor para o seu aprendizado. Para escolher as táticas afetivas adequadas, o agente estuda as emoções do aluno a partir do comportamento observável do aluno, isto é, das ações do aluno na interface do sistema educacional.

3.4 Diferencial do presente trabalho

Os trabalhos encontrados no levantamento bibliográfico tratam diferentes estratégias para o acompanhamento do aprendizado e para capturar o comportamento do aluno como suporte a ambiente EaD.

O diferencial deste trabalho está na sua proposta para desenvolver uma ferramenta que agregue valor qualitativo na interatividade do aluno com o ambiente. Pretende-se assistir

os alunos nas dúvidas durante o processo de aprendizagem e ajudá-lo a melhor entender os conteúdos. Acreditamos que com a ferramenta adequada, o aluno vai certamente achar a motivação para aumentar sua interatividade com o ambiente, e desta forma, a avaliação do comportamento, do nível de conhecimento e do grau de interesse do aluno poderá ser mais produtiva.

Nos próximos capítulos serão descritos o desenvolvimento, testes e resultados do trabalho.

4 *Uma proposta para apoiar a aprendizagem em EaD*

A proposta deste trabalho é apoiar os alunos na sua interação em ambientes de EaD. Em particular, auxiliar o estudante na busca por respostas a suas dúvidas sobre o conteúdo dos cursos disponibilizados no ambiente de EaD.

O objetivo da ferramenta é integrar a busca (da parte dos alunos) e descoberta (da parte da ferramenta) de informação para promover, assim, a comunicação aluno-ambiente. A ferramenta oferece para os alunos suporte para que possam expor, em linguagem natural, suas dúvidas nas diferentes disciplinas do curso de EaD. As perguntas serão respondidas com a melhor resposta possível, a partir da utilização de técnicas de mineração em bases de conhecimento, preenchidas pelos professores especialistas das disciplinas. A interface da ferramenta no ambiente de aprendizagem pode ficar disponível em qualquer um dos mecanismos de comunicação (fóruns, bate-papo, listas de discussão, dentre outros).

Na busca por soluções, foram desenvolvidas duas abordagens que podem se inseridas em um ambiente de EaD. A primeira utiliza a mineração de dados em bases de conhecimento. A segunda usa a mineração de texto. O objetivo final das duas abordagens é apresentar ao aluno a melhor resposta encontrada para sua pergunta.

A seguir são descritas cada uma delas junto com sua arquitetura de funcionamento.

4.1 Utilizando mineração de dados

Nesta primeira abordagem, para alcançar seu objetivo foi utilizada a mineração de dados para classificar uma base de conhecimento que contém as respostas cadastradas pelos professores (neste contexto, atuando como especialistas). A base de conhecimento possui, além das respostas, cinco palavras chaves associadas, relevantes àquele assunto.

A Figura 1 mostra a arquitetura da ferramenta em três camadas. De maneira geral,

as camadas têm os seguintes objetivos:

1. *Camada de Acesso aos Dados*: a tarefa desta camada é acessar e preparar a fonte de dados para a camada de inteligência. A ferramenta WEKA (*Waikato Enviroment for Knowledge Analysis*) (17) tem um formato próprio para especificação dos dados. Como parte desta camada foi desenvolvido um módulo de migração que: lê a base de dados e retorna um conjunto; recebe este conjunto e grava as instâncias na saída padrão, no caso, o arquivo WEKA que será utilizado na camada de inteligência.

2. *Camada de Inteligência*: esta camada tem o objetivo de fornecer o suporte inteligente dentro do ambiente. Para isto, utiliza a técnica de mineração de dados que mais se qualifica para um problema dado em associação com a ferramenta de mineração de dados WEKA. Nesse momento, a técnica é utilizada para organizar os dados da base, por exemplo, os dados podem ser classificados de acordo com um algoritmo de árvore de decisão. Esta estrutura será utilizada mais tarde pela camada de interface.

3. *Camada de Interface*: esta camada serve de interface entre a ferramenta e a aplicação que está fazendo uso dela.

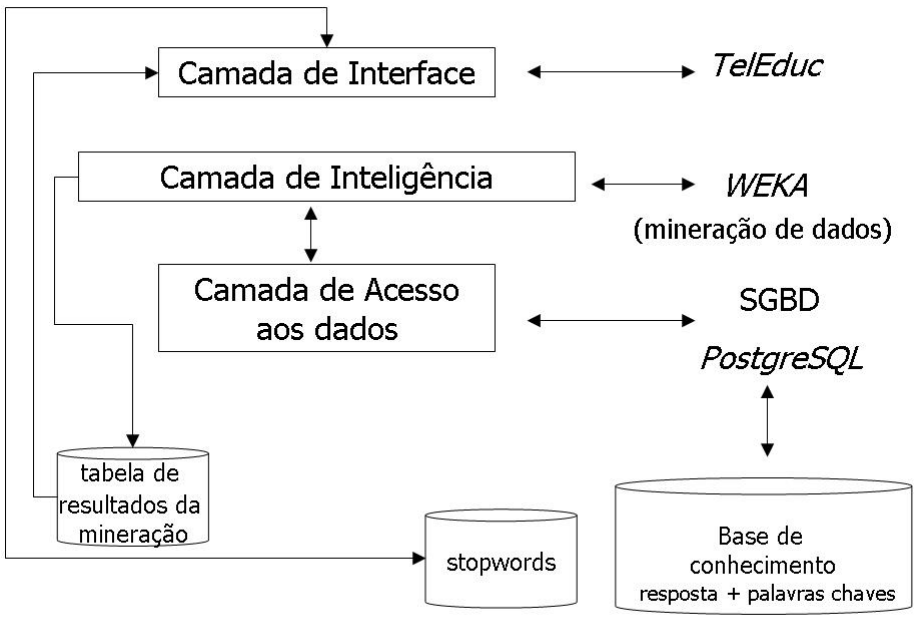

Figura 11: Arquitetura da ferramenta

Em um primeiro momento, a ferramenta irá assistir o aluno respondendo com a resposta ou dúvida melhor classificada pelo algoritmo proposto. Para alcançar esse objetivo foi utilizada a mineração de dados para classificar uma base de conhecimento que contém as respostas cadastradas pelos professores (neste contexto, atuando como especialistas). A base de conhecimento possui, além das respostas, cinco palavras chave associadas, relevantes àquele tópico. Além da base de conhecimento, uma base auxiliar é criada para armazenar as palavras descartáveis ou *stopwords* que ocorrem freqüentemente em textos, tais como artigos, preposições, pronomes, dentre outras. Esta segunda base irá oferecer suporte à análise da pergunta em linguagem natural.

A ferramenta foi desenvolvida sob o ambiente TelEduc. A primeira e a segunda camada foram desenvolvidas em Java 5.0 (1.5.0.03) por ser livre, orientada a objeto e altamente portável, a terceira camada foi desenvolvida em PHP (*PHP: Hypertext Preprocessor*) por haver a necessidade de ficar integrada ao ambiente EaD escolhido.

4.1.1 Camada de acesso aos dados

A camada de acesso a dados utiliza os *drivers* disponíveis da linguagem Java para acesso a diversos sistemas de bancos de dados e serve de interface de comunicação entre as consultas e seus resultados que serão utilizados pela camada de inteligência. A base de conhecimento foi armazenada no sistema de gerenciamento de bancos de dados PostgreSQL.

4.1.2 Camada de inteligência

A Figura 2 mostra a camada de inteligência. Após estudar e testar vários algoritmos de mineração implementados no WEKA baseadas em técnicas como classificação, redes bayesianas e análise de agrupamento. A árvore de decisão foi a técnica que mais se qualificou para a nossa abordagem. A ferramenta disponibiliza diferentes algoritmos (`ID3`, `J48`, `NBTree`, etc.), como mostra a Figura 2.

Esta camada tem o objetivo de montar uma árvore de decisão que particiona recursivamente um conjunto de treinamento, até que cada subconjunto deste particionamento contenha casos de uma única classe. Existem diversas medidas de diversidade envolvendo as densidades de distribuição dos dados em cada classe, geradas pelo teste naquele nó. O ganho de informação, baseado em entropia, tem prevalecido como fato de escolha do atributo a ser testado num nó.

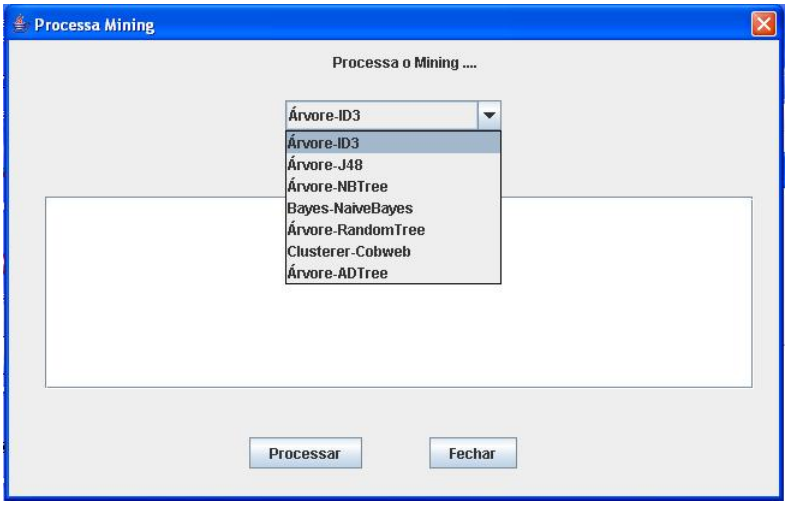

Figura 12: Camada de Inteligência da Ferramenta

O resultado da aplicação do algoritmo é apresentado na tela, possibilitando ao especialista validar o resultado da técnica de mineração de dados escolhida. Confirmada sua escolha, este resultado é cadastrado em uma tabela que representa o resultado da classificação. Esta tabela será utilizada mais tarde na camada de *interface* para pesquisar as respostas às dúvidas dos alunos. Este recurso possibilita o isolamento entre a camada de inteligência e a de *interface*.

4.1.3 Conclusões da utilização de mineração de dados

A utilização da mineração de dados mostrou-se eficiente como um mecanismo para auxiliar o algoritmo de recuperação de respostas. Porém, nos testes realizados ficou demonstrado a vulnerabilidade da ferramenta quanto à correta ordem das palavras chaves na base. Isto é conseqüência do processo de inserção das palavras chaves, que é um processo subjetivo, que depende da ordem em que o professor especialista especifica na hora de inserir as palavras-chaves. A seguir, apresenta-se um exemplo: *o documento a ser cadastrado na base de conhecimento descreve a definição de classe.* A ordem das palavras-chaves é, segundo esperado pela ferramenta, da mais genérica à mais específica:

- palavra-chave1
- palavra-chave2
- palavra-chave3
- palavra-chave4

No entanto, o especialista poderia definir a seguinte ordem:

- palavra-chave4
- palavra-chave2
- palavra-chave3
- palavra-chave1

Esta situação resultaria em problemas de classificação, pois a raiz da arvore deve conter a palavra mais genérica (isto é, espera-se que seja a palavra-chave1). Neste exemplo, a raiz não teria a palavra mais genérica. Assim, a ferramenta fica dependente do critério do especialista. No entanto, se o número de erros de este tipo for pequeno em relação ao tamanho da base de conhecimento não haverá problemas. Em outro caso, esse tipo de situação irá representar um problema na eficácia da ferramenta.

Os testes realizados indicaram que o numero de documentos com erros influenciou o suficiente para gerar a árvore de decisão errada. Os bons resultados, apresentados no próximo capítulo, foram obtidos com esta situação manualmente controlada, tomando-se o devido cuidado no cadastramento das respostas. Esta vulnerabilidade motivou o estudo de outras técnicas para recuperação de informação da base de conhecimento, das quais escolheu-se a mineração de texto.

4.2 Utilizando mineração de texto

Na presente abordagem, as palavras chaves são descobertas associando pesos para cada palavra no texto. A técnica utilizada será a recuperação de informação, focando no modo de descoberta reativa, uma vez que se assume que a base de conhecimento é especifica para o curso em que o aluno está inserido. A base de conhecimento é mantida pelo professor especialista da disciplina e é formada por textos referentes a essa matéria.

Uma vez que a ferramenta de MD não obteve os resultados esperados, foi elaborada uma pesquisa por soluções alternativas que pudessem fornecer melhores respostas. Estas pesquisas nos levaram a estudar mineração de texto.

Com a utilização da MT foi possível descartar a necessidade do especialista (neste cenário, o professor) saber classificar ou fazer qualquer outra operação com as palavras chaves. Estas palavras chaves, com a utilização do MT, são descobertas utilizando-se para

isso a quantidade de vezes que cada palavra é encontrada no texto, criando um mecanismo de peso para cada palavra no texto. O mecanismo utilizado será o RI, mais precisamente o modo de descoberta reativa, uma vez que a base de conhecimento está associado ao curso que o aluno está inserido. Este modo de descoberta foi o escolhido pelos seguintes motivos:

- a base de conhecimento é formada por assuntos específicos;
- o aluno entende o domínio da resposta da pergunta feita, isto é, as respostas obtidas podem ser analisadas pelo aluno;
- há conhecimento do interesse do aluno, definido na pergunta realizada, não havendo a necessidade de descobrir, por exemplo, o que há de relevante em um documento específico, tarefa típica do modo de descoberta pro-ativa;
- não há o risco de sobrecarga de informações uma vez que a base de conhecimento é formada por assuntos específicos.

Desta forma as perguntas serão respondidas com a melhor resposta possível, a partir da utilização de técnicas para minerar textos. A interface da ferramenta no ambiente de aprendizagem pode ficar disponível em qualquer um dos mecanismos de comunicação (fóruns, bate-papo, listas de discussão, dentre outros).

A Figura 3 mostra a arquitetura da ferramenta em três módulos. De maneira geral, os módulos têm os seguintes objetivos:

1. *Módulo de Manutenção*: possibilitar a inclusão, alteração e exclusão de documentos da base de conhecimento.
2. *Módulo de Inteligência*: aplicar a mineração de texto gerando para cada documento a relação de palavras e a quantidade de vezes que a mesma aparece no texto. Esta estrutura será utilizada no mecanismo de recuperação de informação.
3. *Módulo de Interface*: este módulo serve de interface entre a ferramenta e a aplicação que está fazendo uso dela.

A ferramenta irá assistir o aluno com a resposta a sua dúvida melhor classificada pelo algoritmo proposto. Para alcançar tal objetivo foi utilizada técnicas de mineração de textos que classificam uma base de conhecimento. A base de conhecimento contém um

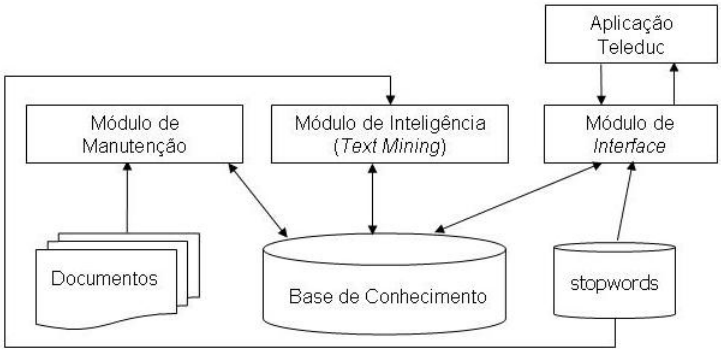

Figura 13: Arquitetura da ferramenta

conjunto de textos cadastrados pelos professores (neste contexto, atuando como especialistas das disciplinas). Uma base auxiliar é criada para armazenar as palavras descartáveis ou *stopwords* que ocorrem freqüentemente em textos, tais como artigos, preposições, pronomes, dentre outras, em português. Esta segunda base irá dar suporte ao processamento da linguagem natural utilizada tanto nos textos das respostas armazenadas na base de conhecimento quanto nas perguntas feitas pelos alunos.

A ferramenta foi desenvolvidos na linguagem PHP 5 (*PHP: Hypertext Preprocessor*), utilizando o SGBD Mysql 5.0 por haver a necessidade de ficar integrado ao ambiente TelEduc, escolhido para esta implementação. Desta forma a interface pode ser integrada a qualquer ambiente de EaD que possua integração à linguagem PHP.

4.2.1 Módulo de manutenção

Este módulo fornece funcionalidades que permitem a administração da base de conhecimento. Para executar a mineração de textos foram feitos vários testes sobre conjuntos de documentos que poderiam vir a formar a base de conhecimento. Por exemplo, estudaram-se as leituras de arquivos nos formatos PDF, DOC, RTF e HTML. Os dois primeiros formatos possuem a desvantagem de serem padrões pagos e, portanto, não poderiam ser disponibilizados num contexto de software livre. Os dois últimos, embora livres, não justificariam sua inclusão na base de conhecimento, devido a sua pouca utilização em materiais didáticos. Desta forma, a base de conhecimento é construída a partir de textos. A *interface* que possibilita este trabalho pode ser vista na Figura 4 .

É importante chamar atenção de como é fundamental incluir conteúdo correto e de

forma correta na base de conhecimento. Qualquer algoritmo de mineração de dados é sensível (não poderia deixar de ser) ao conteúdo no qual está inserido. Portanto, para o correto funcionamento do sistema deve haver uma consciência por parte dos especialistas da necessidade de utilizar e manter corretamente a ferramenta e o conteúdo da base.

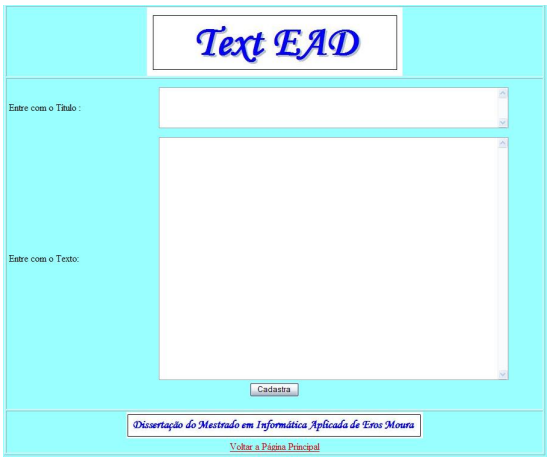

Figura 14: Tela de Cadastro na Base de Conhecimento

4.2.1.1 Consulta, alteração e exclusão da base de conhecimento

A ferramenta disponibiliza funções para verificar quais são os documentos existentes na base de conhecimento e, se necessário, alterar ou excluir algum deles, tornando a possibilidade deste conhecimento estar o mais atualizado possível. As possíveis modificações sobre a base de conhecimento estão associadas a prioridades de usuários que garantem o tipo de acesso aos dados.

4.2.2 Módulo de Inteligência

Neste módulo, os documentos da base de conhecimento serão processados para permitir a recuperação da melhor resposta possível, sempre que o conteúdo da base e a pergunta estejam no idioma português.

1. Limpeza dos índices: no caso de existência de índices antigos é feita uma limpeza na estrutura, pois a cada processo de mineração de texto é feita uma releitura de toda a base de conhecimento. O tempo de processamento está diretamente relacionado ao tamanho da base de conhecimento anterior e configuração de hardware/software que se está utilizando.

2. Leitura das *stopwords*: neste passo ocorre a leitura de todos os documentos existentes na base de conhecimento. O tempo para este processamento está diretamente ligado ao tamanho da base de conhecimento e da configuração de hardware/software que se está utilizando.

3. Leitura dos documentos: neste passo ocorre a leitura de todos os documentos existentes na base de conhecimento. O tempo para este processamento está diretamente ligado ao tamanho da base de conhecimento e da configuração de hardware/software que se está utilizando. Para cada documento:

4. Limpeza de pontos e símbolos: o documento é varrido à procura de caracteres que estão na faixa de 1 a 31 e de 127 a 191, além da procura pelos símbolos, por exemplo :

   ```
   \ . ? ; * ( " ) - < > = + / % | & ^ ~
   ```

5. Limpeza do texto: já com o documento sem pontuação e outros símbolos, todas as *stopwords* encontradas no texto são substituídas pelo caractere espaço. É importante ressaltar a importância de uma boa lista de *stopwords*, pois do contrário, palavras sem significado para pesquisa serão indexadas.

6. Indexação do texto: este é o passo final do algoritmo. Neste ponto, o documento tornou-se um conjunto finito de palavras com significado para nosso contexto (a pergunta feita pelo aluno). Para realizar a indexação, (1) a primeira ação é verificar quantas vezes a palavra aparece em um determinado documento, incluindo tanto o título quanto o próprio texto. (2) Completada a estatística do número de ocorrências das palavras, o índice é criado.

É importante destacar que o algoritmo determina o número de ocorrências das palavras por documento, possibilitando assim uma análise muito mais precisa no momento da pesquisa. Ao completar o processamento, a ferramenta apresenta na tela as estatísticas do número de documentos analisados, número de palavras indexadas no título e número de palavras indexadas no texto.

4.2.3 O módulo de *Interface*

Este módulo pode ser dividido em duas fases:

1. Integração ao TelEduc: como mostra a Figura 5, permite ao aluno expor sua dúvida em linguagem natural (português). Para isto, uma caixa de texto é inserida no ambiente de *chat* do TelEduc. Por exemplo: "O que é objeto".

Figura 15: Página do bate-papo do TelEduc

2. Processamento da pergunta do aluno: esta segunda fase processa a pergunta acompanhando os seguintes passos.

   (a) Limpeza de pontos e símbolos: varre a pergunta procurando caracteres na faixa de 1 a 31 e de 127 a 191, além da procura pelos símbolos.

   (b) Limpeza do texto: já com a pergunta sem pontuação nem símbolos, as *stopwords* são substituídas pelo caractere espaço em branco. É importante ressaltar a importância de uma boa lista de *stopwords*, pois do contrário, palavras sem significado para pesquisa serão utilizadas na recuperação das respostas.

   (c) Recuperação das respostas: a seleção dos documentos que serão retornados é um ponto fundamental para o sucesso do trabalho, mas que depende, como já foi mencionado, da entrada de dados, a lista de *stopwords* e uma pergunta razoavelmente bem elaborada. O algoritmo de seleção (1) para cada palavra selecionada na pergunta é feita uma seleção (2) verifica quais são os documentos que possuem a maior número das palavras selecionadas da pergunta feita pelo usuário. (3) tenta encontrar a resposta, utilizando a conjunção de todos as palavras selecionadas.(4) seleciona os três documentos com maior numero totais de palavras, selecionando os com maior grau de pertinência. (5) Se não

for possível, fará a retirada das palavras, começando pela palavra com menor ocorrência na base e terminando naquela com maior ocorrência, até encontrar a resposta. (5) Se chegando à última, se a resposta não foi encontrada, o algoritmo retornará a falha na busca e na tela será mostrado que não foi possível encontrar uma resposta para a pergunta feita. As respostas retornadas à pergunta mostrada na Figura 5 é apresentada na Figura 6.

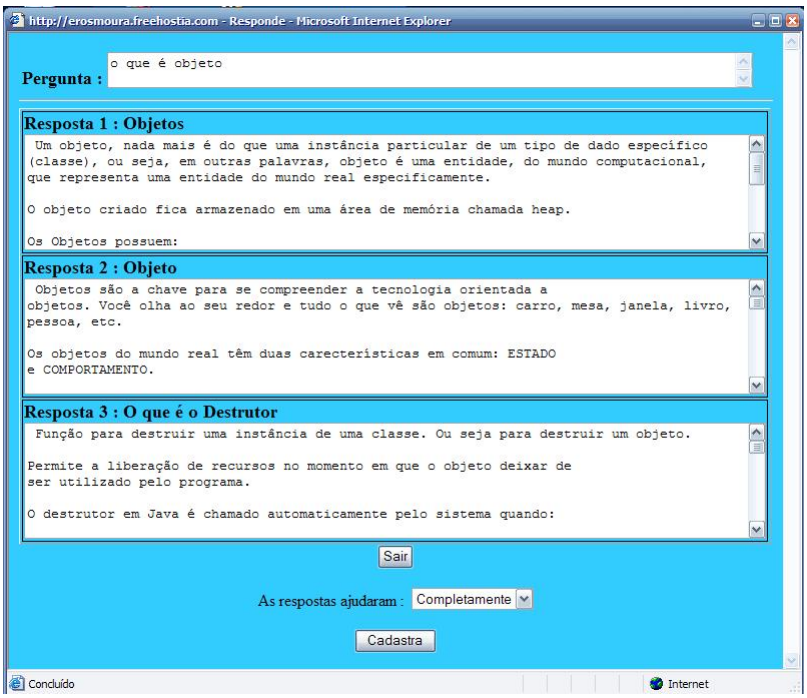

Figura 16: Resposta da ferramenta

Um fator importante no processamento é o tempo de resposta. De nada adiantaria um algoritmo que retornasse os melhores documentos em um tempo maior do que o aceitável para a situação. Como estamos falando de um ambiente virtual *on-line*, obter uma resposta após um minuto de processamento seria inaceitável.

4.3 Problemas e soluções para a ferramenta Text EaD

No processo de mineração de texto podem-se utilizar algoritmos de mineração de dados, isso ocorre sempre em que uma das técnicas escolhida (associação, classificação, etc) traz algum benefício comprovado ao nível de resposta.

Foram feitos vários ensaios, principalmente utilizando algoritmos de classificação, em especial de árvore de decisão (ID3 e C4.5) e qualquer um deles apresentava problemas

devido as características desta aplicação.

Por exemplo, o número estimado para esta base de conhecimento, mesmo depois de algum tempo, não deve chegar a um milhão de documentos, numero que não permitiria aos algoritmos de mineração de dados um resultado satisfatório.

De outro lado, os mecanismos de mineração de dados exigiriam maior tempo de processamento e configuração de máquina e software mais robustos.

Dada a complexidade dos algoritmos utilizados na mineração de texto, foi possível obter respostas satisfatórias no escopo do trabalho, o qual será melhor descriminado no capítulo de estudo de caso.

5 Resultados e Testes

Com o objetivo de validar a eficácia da ferramenta, utilizamos uma instalação do TELEDUC. Esta instalação foi feita no Centro Universitário São Camilo - ES, e descrevemos os resultados obtidos nesta fase utilizando as duas abordagens utilizadas, mineração de dados e mineração de texto.

5.1 O ambiente para a Mineração de Dados

O ambiente criado é formado por uma máquina Pentium 4 2.8 com 256 Mb de memória e 80 Gb de *hard disk*. Foi utilizada a distribuição do Linux chamada de Kurumin, com PHP 5, Mysql 4.0.3 e Apache 2.0.2. Este ambiente tem sido utilizado como testes para estudos sobre EaD e a participação dos alunos é opcional, através de convite, para complementar conteúdos não abordados de forma convencional em sala de aula.

Uma base de conhecimento foi montada com 580 respostas, cadastradas pelos professores (neste contexto, atuando como especialistas). Esse conjunto de respostas teve como domínio quatro disciplinas na área de Sistemas de Informação, como mostra a tabela **??**.

Assunto Principal	**Quantidade**
Programação Orienta a Objeto	85
Linguagem de Programação Java	245
Linguagem de Programação PHP	130
Banco de Dados Relacional	120

Tabela 9: Classificação das respostas cadastradas

Cada uma das respostas cadastradas pode ter de uma a cinco palavras chave. É baseado neste conjunto de palavras chave que será aplicado um algoritmo de classificação da mineração de dados.

Montada a base de conhecimento, foi feita a seleção, preparação e limpeza dos dados. Em seguida, os dados foram recuperados e submetidos aos algoritmos disponíveis na

ferramenta. Após 230 simulações de conjuntos de palavras chave, chegou-se a conclusão que o algoritmo que apresentou melhores resultados foi o `C4.5`, implementado pelo método `J48`. O algoritmo `C4.5` pode ser estudado em (**??**).

A árvore de decisão resultante foi armazenada numa tabela de resultados. Nesta árvore, os atributos que aparecem nos primeiros nodos são os atributos de maior ganho de informação.

5.1.1 Camada de *interface*

A utilização da mineração de dados mais o algoritmo implementado na camada de *interface* possibilita encontrar o melhor conjunto de respostas existente na base de conhecimento, caso exista. Parte do código foi inserido no TelEduc, mais precisamente no código da janela de bate-papo (*chat*). O projeto tentou minimizar as alterações no ambiente TelEduc, visando facilitar a utilização dessa ferramenta pelo seu grupo de usuários. Logo, o aluno passa a ter a sua disposição um campo na janela que pode digitar uma pergunta qualquer.

5.1.2 Aplicando a ferramenta

Na janela de bate-papo do TelEduc (Fig. **??**), o usuário escreve a pergunta, utilizando linguagem natural e clica em *Fazer Pergunta à Base de Conhecimento*.

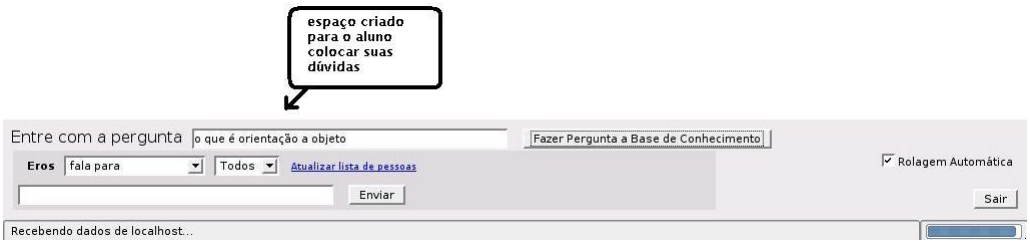

Figura 17: Página do bate-papo do Teleduc

Ao se digitar uma pergunta, a ferramenta executa o seguinte algoritmo de pesquisa e recuperação de respostas:

1. A pergunta é processada. Para fazer a análise da pergunta em linguagem natural, descartam-se aquelas palavras encontradas na tabela das *stopwords*. As palavras remanescentes são chamadas de palavras chave à consulta e são utilizadas na pesquisa.

2. Uma consulta é construída como a conjunção de restrições e baseada no resultado da classificação armazenada na tabela de resultados da mineração de dados. As palavras chave representam os valores sobre os atributos da árvore, mapeadas uma a uma a partir da raiz.

3. Executa esta consulta. Caso não encontre resultados, tenta obter respostas retirando palavras chave da consulta até encontrar um conjunto não vazio ou chegar à raiz. Isto é feito subtraindo da restrição a última condição, o que significa subir um nível na árvore.

4. Retorna para o aluno a melhor resposta encontrada ou, caso não encontre, a informação de que não possui resposta para a pergunta feita.

As respostas resultantes são visualizadas na tela, como mostra a Figura **??**.

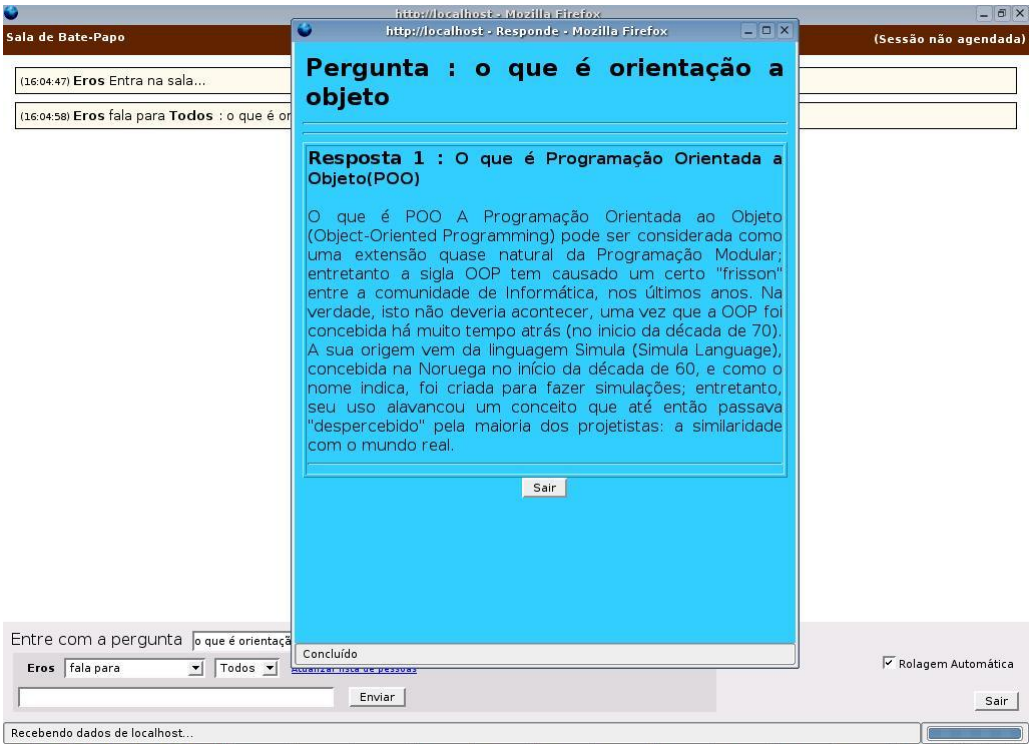

Figura 18: Resposta à dúvida apresentada pelo aluno

5.1.3 Resultados com a Mineração de Dados

Nos testes realizados, com o objetivo de validar a ferramenta, foi utilizado um conjunto de perguntas que podem ser classificadas utilizando-se o critério do assunto principal, mostrado na tabela **??**.

	Assunto Principal	Quantidade
1	Programação Orienta a Objeto	45
2	Linguagem de Programação Java	35
3	Linguagem de Programação PHP	20
4	Banco de Dados Relacional	15

Tabela 10: Classificação das perguntas realizadas

Como estudo de caso foi feito um teste com 115 (cento e quinze) perguntas. Dentre elas: *o que é um objeto* (associada ao tópico 1) , *o que é integridade de dados* (associada ao tópico 4), *o que é método* (associada ao tópico 2), etc.

O objetivo era simular um ambiente real com perguntas, das mais variadas formas. A resposta foi a esperada em 75% das perguntas. Os 25% não satisfatórios foram atribuídos à qualidade da base de conhecimento, pois se verificou que para o correto funcionamento do algoritmo de classificação na mineração de dados há a necessidade da correta disposição das palavras chave, de tal forma que a palavra mais abrangente venha sempre antes de uma mais específica. Isto nem sempre é viável, pois é preciso contar com a subjetividade do especialista na sua escolha e avaliação das palavras chave.

5.2 O ambiente para a Mineração de Texto

O ambiente criado é formado por uma máquina Pentium 4 2.8 com 256 Mb de memória e 80 Gb de hard disk.

Foi utilizada a distribuição do Linux chamada de Kurumin, com PHP 5, Mysql 4.0.3 e Apache 2.0.2.

Este ambiente tem sido utilizado como testes para estudos sobre EaD e a participação dos alunos é opcional, através de convite, para complementar conteúdos não abordados de forma convencional em sala de aula.

A base de conhecimento foi montada para as áreas de Ciência da Computação. A Tabela **??** descreve seu conteúdo.

	Assunto Principal	Quantidade
1	Programação Orientada a Objeto	45
2	Linguagem de Programação Java	35

Tabela 11: Classificação das áreas de conhecimento

5.2.1 Resultados com a Mineração de Texto

Com o fim de validação, a ferramenta foi implantada na Universidade Cândido Mendes (campus Campos dos Goytacazes) e no Centro Universitário São Camilo (ES). Os alunos dos cursos de Ciência da Computação e Sistemas de Informação fizeram os testes da ferramenta. O objetivo foi simular um ambiente real com perguntas, das mais variadas formas.

Além da pergunta, o aluno podia especificar o grau de satisfação com a resposta retornada pela ferramenta, classificado em *bom*, *regular* e *ruim*. Uma tabela foi criada para armazenar as perguntas feitas pelos alunos e o grau de satisfação. Isto permitiu descartar aquelas perguntas cujo conteúdo não encontrava-se incluído na base de conhecimento.

Os testes chegaram completar 150(cento e cinqüenta) perguntas. Dentre elas: *o que é objeto* (associada ao tópico 1), *como declarar uma classe em Java* (associada ao tópico 2), etc.

A resposta foi satisfatória em 85,33% das perguntas, sendo que este cálculo exclui as perguntas descartadas. Em 3,33% das perguntas efetuadas obteve grau de satisfação regular e 11,34% obteve avaliação ruim. Os casos não satisfatórios poderiam ser atribuídos à qualidade da base de conhecimento, pois verificou-se que alguns textos cadastrados na base de conhecimento eram muito abrangentes e muito longos. O tempo de resposta às perguntas foi considerado satisfatório (em média dois segundos).

6 Conclusões e Trabalhos Futuros

O ambiente de EaD é um espaço onde a tecnologia pode ajudar muito para melhorar o aprendizado do aluno. Uma tentativa interessante foi feita no nosso trabalho para melhorar a comunicação aluno-ambiente, tentando auxiliar ao aluno na busca de respostas a suas dúvidas. O estudo tentou minimizar as alterações no ambiente TelEduc, visando facilitar a utilização dessa ferramenta pelo seu grupo de usuários.

A utilização da mineração de dados mostrou-se eficiente como um mecanismo para auxiliar o algoritmo de recuperação de respostas. Porém, nos testes realizados ficou demonstrado a vulnerabilidade da ferramenta quanto à correta ordenação das palavras-chaves pelo especialista ao inserir na base de dados.

Esta vulnerabilidade motivou pesquisas na utilização de outras técnicas para recuperação da base de conhecimento, em especial, mineração de texto.

A utilização da mineração de texto mostrou-se eficiente como um mecanismo para auxiliar a pesquisa e recuperação de respostas para os alunos. Porém, nos testes realizados ficou demonstrada a vulnerabilidade da ferramenta quanto ao texto inserido na base de conhecimento. O texto deve ser o suficientemente claro e objetivo para tratar de um determinado tema. O controle e a prevenção de erros devem ser feitos pelo próprio especialista, que alimenta a base de conhecimento, cuja seriedade na realização desse trabalho deve prevalecer.

Pretende-se, também, que a ferramenta inclua técnicas de mineração de texto sobre bases históricas não estruturadas que contenham informações sobre os alunos e seu comportamento. Com o uso do computador, tornou-se possível capturar algumas características do aprendiz à distância, e analisá-las de uma maneira análoga ao comportamento de um aluno de um curso presencial. A linguagem corporal, o grau de interesse, a participação, o comportamento social, podem ser vistos pela ótica computacional, considerando, basicamente, as interações do aluno com o ambiente de ensino. A freqüência de sua participação em listas de discussão, conferências, fórum, *e-mails* e *chats*, por exem-

plo, podem retratar sua sociabilidade. Outros pontos que poderiam ser objeto de novos estudos seriam: a utilização de outros formatos de texto e a elaboração de novos testes com a utilização de uma base de conhecimento com um volume maior de documentos.

Referências

1 SILVA, D. R. *Um Ambiente para Descoberta de Conhecimento com Suporte de Data Warehousing e sua Aplicação para Acompanhamento do Aluno em Educação a Distância*. Dissertação (Mestrado) — UFSCAR, 2002.

2 JOHNSON, J. Data-driven school improvement. In: *ERIC Clearinghouse on Educational Management Eugene*. [S.l.]: ERIC Digest, 2000. v. 109.

3 MEHLECKE, Q. T.; TAROUCO, L. M. R. Ambientes de suporte para educaÇÃo a distÂncia:a mediação para aprendizagem cooperativa. *CINTED-UFRGS*, Abril, 2003.

4 PETERS, O. Didática do ensino a distância. são leopoldo,rs -unisinos. *Revista Brasileira de Aprendizagem Aberta e a Distância*, 2001.

5 LANDIM, C. M. *Educação à Distância: Algumas Considerações*. [S.l.]: Cibercultura, 1999.

6 GAVA, T. *Estações de Aprenizagem: um modelo baseado em ontologias - UFES*. Dissertação (Mestrado) — UFES, 2003.

7 PERRONE, J. *EDUCNET: AMBIENTE INTERATIVO PARA EDUCAÇÃO A DISTÂNCIA ON-LINE*. Dissertação (Mestrado) — Fundação Visconde de Cairu - Salvador - Bahia, 2005.

8 ROCHAM, H. et al. Design de ambientes para ead: (re)significações do usuário. *Anais do IHC´2001 - IV Workshop sobre Fatores Humanos em Sistemas Computacionais - Florianópolis, SC, pg 84-9*, 2001.

9 LOPES, C. C. *Um sistema de apoio à tomada de decisão no acompanhamento do aprendizado em educação a distância*. Dissertação (Mestrado) — UFCG, 2003.

10 MACHADO, L.; BECKER, K. O uso da mineração de dados na web aplicado a um ambiente de ensino a distância. *I Workshop de Teses e Dissertações em Banco de Dados, XIX Simpósio Brasileiro de Banco de Dados*, Gramado, 2002.

11 OLIVEIRA, A. G.; GARCIA, D. F. Mineração da base de dados de um processo seletivo universitário. *INFOCOMP - Journal of Computer Science*, v. 3, n. 2, p. 38–43, 2004. DCC, Federal University of Lavras, Brazil.

12 OEIRAS, J. Y. Acel - ambiente computacional auxiliar ao ensino/aprendizagem a distância de línguas. *UNICAMP*, 1998.

13 JAQUES, P. A.; OLIVEIRA, F. Agentes de software para análise das interações em um ambiente de ensino a distância. In: *III InfoEducar*. Fortaleza, Brasil: [s.n.], 1998.

14 MENEZES, R. A.; FUKS, H.; GARCIA, A. B. Utilizando agentes no suporte à avaliação informal no ambiente de instrução baseada na web aulanet. In: *X Simpósio de Brasileiro de Informática na Educação*. Fortaleza, Brasil: [s.n.], 1999.

15 GUEDES, G. T.; VICCARI, R. M.; DAMICO, C. B. Uma ferramenta para auxiliar a avaliação de textos construídos colaborativamente em ambientes de ensino-aprendizagem. *Revista do CCEI (Centro de Ciências da Economia e Informática), URCAMP*, v. 6, n. 9, 2002.

16 JAQUES, P. A.; VICCARI, R. M. Pat: um agente pedagógico animado para interagir afetivamente com o aluno. *Novas Tecnologias na Educação CINTED-UFRGS*, v. 3, n. 1, 2005.

17 WITTEN, I.; FRANK, E. *Data Mining: practical machine learning tools and techniques with Java implementations*. [S.l.]: Morgan Kaufmann, 2005.

ANEXO A – Relação dos códigos fonte da ferramenta Text EaD

Relação dos programas que implementam mineração de texto como mecanismo de acesso aos dados. Estes programas foram desenvolvidos na linguagem PHP, utilizando como banco de dados o Mysql.

Neste anexo temos os códigos da ferramenta Text EaD.

A.1 Programa principal do Text EaD

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html><!-- InstanceBegin template="/Templates/li_modelo.dwt"
codeOutsideHTMLIsLocked="false" --> <head> <!--
InstanceBeginEditable name="doctitle" --> <title>TextEAD</title>
<!-- InstanceEndEditable --> <meta http-equiv="Content-Type"
content="text/html; charset=iso-8859-1"> <!--
InstanceBeginEditable name="head" --><!-- InstanceEndEditable -->
</head>

<body bgcolor="#99FFFF" text="#000000" link="#333399"
vlink="#CC0000" alink="#663399" topmargin="1"> <table width="899"
border="1" align="center">
  <tr>
    <td height="95" align="center"><div align="center">
    <img src="imagens/logo.JPG" width="425" height="114" border="0">
    </div></td>
  </tr>
  <tr> <!-- InstanceBeginEditable name="conteudo" -->
    <td align="center"> <p><a href="grava_bc.htm">
    Incluir na Base de Conhecimento</a></p>
      <p><a href="./php/escolhe_bc.php">
      Visualizar/Alterar/Excuir na Base de Conhecimento</a></p>
      <p><a href="php/textmining.php">
      Criar os &Iacute;ndices para o Text Mining</a></p>
      <p><a href="php/entrada_dados.php">
      Fazer Pergunta a Base de Conhecimento</a></p>
      <p> </p></td>
```

```
    <!-- InstanceEndEditable --></tr>
  <tr>
    <td align="center"> <div align="center">
    <img src="imagens/rodape.JPG" width="493" height="44" border="0">
      </div>
      <a href="textead.html">Voltar a P&aacute;gina Principal</a></td>
  </tr>
</table> </body> <!-- InstanceEnd --></html>


\normalsize
```

A.2 Programa para armazenar os documentos em PHP

```
Programa que permite armazenar os documentos
na base de conhecimento do Text EaD
A parte do HTML

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html><!-- InstanceBegin template="/Templates/li_modelo.dwt"
codeOutsideHTMLIsLocked="false" --> <head> <!--
InstanceBeginEditable name="doctitle" --> <title>Avaliação</title>
<!-- InstanceEndEditable --> <meta http-equiv="Content-Type"
content="text/html; charset=iso-8859-1"> <!--
InstanceBeginEditable name="head" --><!-- InstanceEndEditable -->
</head>

<body bgcolor="#99FFFF" text="#000000" link="#333399"
vlink="#CC0000" alink="#663399" topmargin="1"> <table width="899"
border="1" align="center">
  <tr>
    <td height="95" align="center"><div align="center">
    <img src="imagens/logo.JPG" width="425" height="114"
    border="0"></div></td>
  </tr>
  <tr> <!-- InstanceBeginEditable name="conteudo" -->
    <td align="center" valign="top">
      <form action="./php/cadastra_bc.php" method="post" name="form1" >
        <table width="100%" height="440" border="0">
          <tr>
            <td width="22%" height="90" align="left" valign="middle">
            Entre com o T&iacute;tulo : </td>
            <td width="78%" align="center"
            valign="middle"><textarea name="titulo" cols="70" rows="4"
            id="titulo"></textarea></td>
          </tr>
          <tr>
            <td height="344" align="left"
            valign="middle" bordercolor="#3300FF">Entre
              com o Texto:</td>
            <td align="center" valign="middle">
```

```
            <textarea name="texto" cols="70" rows="25" id="texto">
            </textarea></td>
          </tr>
        </table>
        <input name="Cadastra" type="submit" value="Cadastra">
      </form>
    </td>
    <!-- InstanceEndEditable --></tr>
  <tr>
    <td align="center"> <div align="center"><img
    src="imagens/rodape.JPG" width="493" height="44" border="0">
      </div>
      <a href="textead.html">Voltar a P&aacute;gina Principal</a></td>
  </tr>
</table> </body> <!-- InstanceEnd --></html>

A parte do PHP

<?php
   /* Cadastro na Base de Conhecimento */
   class Grava_BC{

       function conecta() {
          require 'conecta.inc';
       }

       function exec_query($query){
          $rs = mysql_query($query);
          if (!$rs) {
             die("Erro na Execução da Query : ".$query );
          }
          else {
             return $rs;
          }
       }//fim da funcao
   }

   $cd_documento = $_POST['cd_documento'];

   $titulo       = $_POST['titulo'];
   $texto        = $_POST['texto'];

   if (isset($_POST['excluir'])) {
    $excluir = $_POST['excluir'];
   }
   else {
    $excluir = "N";
   }

   $grava_bc = new Grava_BC();

   $grava_bc->conecta();

   require('antes.inc');
```

```
   if ($excluir == "S") {
      $query = "delete from documentos where cd_documento = ".
      $cd_documento;
      $rs  = $grava_bc->exec_query($query);
      if (mysql_affected_rows()==-1){
         echo ("<a href='altera_bc.php?cd_documento=".$cd_documento."'>
         Não foi Excluido na base de Conhecimento!.
         Clique aqui para uma nova Tentativa</a>");
      }
      else {
         echo ("<a href='escolhe_bc.php'>
         Foi Excluido na base de Conhecimento!. Clique aqui</a>");
      }
   }
   else {
      $query = "update documentos set titulo = '$titulo',
      texto = '$texto' where cd_documento = ".$cd_documento;
      $rs  = $grava_bc->exec_query($query);
      if (mysql_affected_rows()==-1){
         echo ("<a href='altera_bc.php?cd_documento=".
         $cd_documento."'>Não foi Alterado na base de Conhecimento!.
         Clique aqui para uma nova Tentativa</a>");
      }
      else {
         echo ("<a href='altera_bc.php?cd_documento=".
         $cd_documento."'>Foi Alterado na base de Conhecimento!.
         Clique aqui</a>");
      }
   }
   require('depois.inc');
?>
```

A.3 Programa para alterar os documentos em PHP

```
Programa que permite alterar os documentos na
base de conhecimento do Text EaD
Parte do PHP

<?php
   /* Cadastro na Base de Conhecimento */
   class Manutencao_BC{

       function conecta() {
          require 'conecta.inc';
       }

       function exec_query($query){
          $rs = mysql_query($query);
          if (!$rs) {
             die("Erro na Execução da Query : ".$query );
```

```
        }
        else {
           return $rs;
        }
    }//fim da funcao
}

$cd_documento = $_GET['cd_documento'];

$manutencao_bc = new Manutencao_BC();

$manutencao_bc->conecta();

$query = "select titulo,texto from documentos where cd_documento = ".
$cd_documento;
$rs  = $manutencao_bc->exec_query($query);
require('antes.inc');

if (mysql_num_rows($rs)==0){
   echo ("<a href='altera_bc.php'>
   Sem dados na base de Conhecimento!.
   Clique aqui para uma nova Tentativa</a>");
}
else {
   $linha = mysql_fetch_array($rs,MYSQL_ASSOC);
 ?>
     <td align="center" valign="top">
       <form action="./grava_bc.php" method="post" name="form1" >
        <input name="cd_documento" type="hidden" value=<?php
        echo($cd_documento)?> >
         <table width="97%" height="524" border="1">
           <tr>
             <td width="25%" height="90">
             Entre com o T&iacute;tulo : </td>
             <td width="75%"><textarea name="titulo" cols="99"
             rows="4" id="titulo"><?php echo($linha['titulo']) ?>
             </textarea></td>
           </tr>
           <tr>
             <td height="426" valign="middle">
             Entre com o Texto:</td>
             <td><textarea name="texto" cols="99" rows="25"
             id="texto"><?php echo($linha['texto']) ?></textarea></td>
           </tr>
           <tr align="center">
             <td colspan="2" valign="top">
            <input name="excluir" type="checkbox" value="S">
            Quer Excluir
             </td>
           </tr>
          <tr align="center">
            <td colspan="2" valign="top">
            <input name="Altera" type="submit" value="Gravar">
            </td>
```

```
            </tr>
          </table>
         </form>
       </td>
     <?php
  }
  require('depois.inc');

?>
```

A.4 Programa escolhe o documento em PHP

```
Programa que permite escolher em qual documento se quer
trabalhar na base de conhecimento do Text EaD
Parte em PHP

<?php
  /* Cadastro na Base de Conhecimento */
  class Escolhe_BC{

      function conecta() {
         require 'conecta.inc';
      }

      function exec_query($query){
         $rs = mysql_query($query);
         if (!$rs) {
            die("Erro na Execução da Query : ".$query );
         }
         else {
            return $rs;
         }
      }//fim da funcao
  }

  $escolhe_bc = new Escolhe_BC();

  $escolhe_bc->conecta();

  $query = "select cd_documento,titulo from documentos ";

  $rs  = $escolhe_bc->exec_query($query);

  require('antes.inc');

  if (mysql_num_rows($rs)==0){
     echo ("<a href='../textead.html'>
     Sem dados na base de Conhecimento!.
     Clique aqui para uma nova Tentativa</a>");
  }
  else {
```

```
      echo("<table width='100%' border='1'>");
      echo("<tr>");
      echo("<td width='10%'>Código</td>");
      echo("<td width='90%'>Título</td>");
      echo("</tr>");
      while ($linha = mysql_fetch_array($rs,MYSQL_ASSOC)){
         echo("<tr>");
         echo("<td width='10%'>"."<a href='./altera_bc.php?
         cd_documento=".
         $linha['cd_documento']."'>".
$linha['cd_documento']."</td>");
         echo("<td width='90%'>".substr($linha['titulo'],0,30)."</td>");
         echo("</tr>");
      }
      echo("</tr>");
      echo("</table>");
   }
   require('depois.inc');

?>
```

A.5 Programa mineração de texto em PHP

```
Programa que implementa a mineração de texto na
base de conhecimento do Text EaD
A parte de PHP

<? $v_totais[0]=0;

class TextMining {
   function conecta() {
      // Conecta a base de Dados
      require_once('conecta.inc');
   }

   function exec_query($query){
      //Executa a query
      $rs = mysql_query($query);
      if (!$rs) {
         die("Query com Erro: ".$query." - ".mysql_error()."<br>");
      }
      else {
         return $rs;
      }
   }

   function carrega_stopwords(){
      //Executa a query
      $rs = $this->exec_query("Select * from stopwords;");
      $i = 0;
```

```
   $v_stopwords[$i]="";
   while ($linha = mysql_fetch_array($rs,MYSQL_ASSOC)){
      $v_stopwords[$i] = $linha['stopword'];
      $i++;
   }
   return($v_stopwords);
}


function troca_pontos($text){
   $retira = "\\\ \, \. \? \; \* \( \" \) \- \< \> \= \+ \/
   \% \| \& \^ \~ \´ \` \' \* \# \@ \$ \: \! 1 2 3 4 5 6 7 8 9 0";
   $pontos = explode(" ", $retira);
   foreach($pontos as $key=>$ponto)
   {
      $text = eregi_replace(trim($ponto),' ',$text);

      for ($i=1;$i<32;$i++) {
        $text = eregi_replace(chr($i),' ',$text);
      }

      for ($i=127;$i<192;$i++) {
        $text = eregi_replace(chr($i),' ',$text);
      }

   }
   return $text;
}

function limpa_texto($text,$v_stopwords){
   $v_text           = explode(" ", $text);
   $j                = 0;
   $v_texto_limpo[$j] = '';

   foreach($v_text as $i=>$palavra)
   {
      // Verifica se a palvra da vez é uma que não serve
      $serve = FALSE;
      foreach($v_stopwords as $key=>$stopword)
      {
         // passar para UPPER para comparar
         if (strtoupper(trim($palavra))==
         strtoupper(trim($stopword))) {
            // Se não servir coloco "" substituindo
            $serve = FALSE;
            break;
         }
         else {
            if (strlen(trim($palavra))>0) {
               // Se não encontrar no vetor palavras_fora é
               //porque serve, então continua
               $serve = TRUE;
            }
         }
```

```
        }
        // Se encontrar uma palvra que serve adiciono a
        //mesma no vetor de retorno
        if ($serve) {
           $v_texto_limpo[$j] =  $palavra;
           $j++;
        }
     }
     return $v_texto_limpo;
  }


  function calcula_quantidade($v_palavras){
  // Método para calcula a quantidade de cada palavra
  //em cada Documento
     $v_q_palavras = array_count_values($v_palavras);
     return($v_q_palavras);
  }

  function armazena_totais($palavra,$cd_documento,
  $q_palavras_titulo,$q_palavras_corpo) {
     $query = "select count(*) as quant
                 from palavras_documento
                 where palavra='".trim($palavra)."'
                   and cd_documento = ".$cd_documento;
     $rs = $this->exec_query($query);
     $linha = mysql_fetch_array($rs);
     if ($linha['quant']>0) {
        /* já existe faço um update */
        $query = "update palavras_documento
                    set quantidade_texto = $q_palavras_corpo
                    where palavra='".trim($palavra)."'
                      and cd_documento = ".$cd_documento;
     }
     else {
        $query = "insert into palavras_documento
                     (palavra,cd_documento,quantidade_titulo,
                      quantidade_texto)
                      values ('".strtolower(trim($palavra))."',".
                      $cd_documento.",".
$q_palavras_titulo.",".$q_palavras_corpo.")";
     }
     $rs = $this->exec_query($query);
     return($rs);
  }

}


$textmining = new TextMining();


/* Conecta a base de dados */ $textmining->conecta(); /* Faz uma
limpeza na tabela */ $query = "delete from palavras_documento";
```

```
$resultado = $textmining->exec_query($query);

/* Carrega um vetor com as stopwords  */ $v_stopwords=
$textmining->carrega_stopwords();

/* Vou calcular a quantidade de palavras em cada documento
(registro) da base de conhecimento (titulo e corpo) */ $query =
"select cd_documento,titulo,texto from documentos"; $rs    =
$textmining->exec_query($query); $quantidade_documentos = 0;
$quantidade_titulo   = 0; $quantidade_palavras = 0;

while ($linha = mysql_fetch_array($rs)) {
   $quantidade_documentos++;
   /* Calcula para o título do documento */
   $cd_documento = $linha['cd_documento'];

   // Tiro pontos, virgulas, etc
   $semponto        =
   $textmining->troca_pontos(trim($linha['titulo']));
   // Monto a frase apenas com as palavras corretas
   $v_texto_correto =
   $textmining->limpa_texto(trim($semponto),$v_stopwords);
   if (count($v_texto_correto)>0) {
      $v_q_palavras =
      $textmining->calcula_quantidade($v_texto_correto);

      foreach($v_q_palavras as $palavra=>$qpalavra)
      {
         $palavra     = strtolower($palavra);
         $query = "insert into palavras_documento
                     (palavra,cd_documento,quantidade_titulo,
                      quantidade_texto)
                      values ('".trim($palavra)."',".
                      $cd_documento.",".$qpalavra.",0)";
         $resultado = $textmining->exec_query($query);
         $quantidade_titulo = $quantidade_titulo + $qpalavra;
      }
   }

   /* Calcula para o corpo do documento */

   // Tiro pontos, virgulas, etc
   $semponto        =
   $textmining->troca_pontos(trim($linha['texto']));
   //echo ("<br>Sem os pontos : ".$semponto."<br><br>");
   // Monto a frase apenas com as palavras corretas
   $v_texto_correto = $
   textmining->limpa_texto(trim($semponto),$v_stopwords);

   if (count($v_texto_correto)>0) {
      $v_q_palavras =
      $textmining->calcula_quantidade($v_texto_correto);
      /* Armazena a palavra neste docuemnto com seu total */
      foreach($v_q_palavras as $palavra=>$qpalavra)
```

```
      {
         $palavra   = strtolower($palavra);
         $resultado = $textmining->armazena_totais($palavra,
         $cd_documento,0,$qpalavra);
         $quantidade_palavras =
         $quantidade_palavras + $qpalavra;
      }
   }
} echo ("Quantidade de Documentos         :
".$quantidade_documentos."<br><br>"); echo ("Quantidade de
Palavras no Titulo : ".$quantidade_titulo."<br><br>"); echo
("Quantidade de Palavras no Texto  :
".$quantidade_palavras."<br><br>"); ?>
```

A.6 Programa retorna os documentos em PHP

```
Programa que implementa o retorno dos documentos
selecionados a partir na base de conhecimento do Text EaD
A parte de PHP


<?php

class Retorna_Resposta {
   function conecta() {
   // Conecta a base de Dados
      require_once('conecta.inc');
   }

   function exec_query($query){
      //Executa a query
      $rs = mysql_query($query);
      if (!$rs) {
         die("Query com Erro: ".$query." - ".mysql_error()."<br>");
      }
      else {
         return $rs;
      }
   }

   function carrega_stopwords(){
      //Executa a query
      $rs = $this->exec_query("Select * from stopwords;");
      $i = 0;
      $v_stopwords[$i]="";
      while ($linha = mysql_fetch_array($rs,MYSQL_ASSOC)){
         $v_stopwords[$i] = $linha['stopword'];
         $i++;
      }
```

```
   return($v_stopwords);
}

function troca_pontos($text){
   $retira = "\\\ \, \. \? \; \* \( \" \) \- \< \> \= \+ \/
   \% \| \& \^ \~ \´ \´ \’ \* \# \@ \$ \: \! 1 2 3 4 5 6 7 8 9 0";
   $pontos = explode(" ", $retira);
   foreach($pontos as $key=>$ponto)
   {
      $text = eregi_replace(trim($ponto),’ ’,$text);

      for ($i=1;$i<32;$i++) {
        $text = eregi_replace(chr($i),’ ’,$text);
      }

      for ($i=127;$i<192;$i++) {
        $text = eregi_replace(chr($i),’ ’,$text);
      }

   }
   return $text;
}

function limpa_texto($text,$v_stopwords){
   $v_text           = explode(" ", $text);
   $j                = 0;
   $v_texto_limpo[$j] = ’’;

   foreach($v_text as $i=>$palavra)
   {
      // Verifica se a palvra da vez é uma que não serve
      $serve = FALSE;
      foreach($v_stopwords as $key=>$stopword)
      {
         // passar para UPPER para comparar
         if (strtoupper(trim($palavra))==
         strtoupper(trim($stopword))) {
            // Se não servir coloco "" substituindo
            $serve = FALSE;
            break;
         }
         else {
            if (strlen(trim($palavra))>0) {
               // Se não encontrar no vetor
               // palavras_fora é porque serve,
               // então continua
               $serve = TRUE;
            }
         }
      }
      // Se encontrar uma palvra que serve adiciono
      // a mesma
      // no vetor de retorno
      if ($serve) {
```

```
         $v_texto_limpo[$j] =  $palavra;
         $j++;
      }
   }
   return $v_texto_limpo;
}

function calcula_mais_palavras($v_frase_correta) {
   foreach($v_frase_correta as $i=>$palavra) {
      $query = "select sum(quantidade_texto) as quant
                  from palavras_documento
                  where palavra='$palavra';" ;
      $rs = $this->exec_query($query);
      $linha = mysql_fetch_array($rs);
     $v_quantidade_palavras[$palavra] = $linha['quant'];
   }
   return($v_quantidade_palavras);
}

function verifica_se_tem($v_mais_palavras) {
 $tem = true;
   foreach($v_mais_palavras as $i=>$totais) {
       if ($totais == 0) {
          $tem = false;
       }
       else {
          $tem = true;
       }
   }
   return($tem);
}



function retorna_docs($v_quantidade_palavras){
   $tamanho_vetor = count($v_quantidade_palavras);
   arsort($v_quantidade_palavras);
   reset($v_quantidade_palavras);
   $num_rows   = 0;
   $total_loop = $tamanho_vetor;

   $query = "select cd_documento,titulo, texto
               from documentos
               where cd_documento in (
               select cd_documento from
                     (select cd_documento,
                      sum(quantidade_texto) quant
                  from palavras_documento
                  where ";
   $i=1;
   foreach($v_quantidade_palavras as $palavra=>$totais) {
      if ($i == 1) {
         $query = $query . " palavra = '".$palavra."' ";
      }
      else {
```

```
            if ($i <= $total_loop) {
               $query = $query . " OR palavra ='".$palavra."' ";
            }
         }
         $i++;
      }
      $query = $query." group
      by cd_documento order by quant DESC LIMIT 3 ) maiores) ";
      $rs       = $this->exec_query($query);
      $num_rows = mysql_num_rows($rs);
      return($rs);
    }

}


$retorna_resposta = new Retorna_Resposta();

/* Conecta a base de dados */
$retorna_resposta->conecta(); /*
Carrega um vetor com as stopwords  */ $v_stopwords=
$retorna_resposta->carrega_stopwords();

$texto           = $_GET['pergunta'];

//$texto           = "O que é orientação a objeto";

// Tiro pontos, virgulas, etc
$semponto        =
$retorna_resposta->troca_pontos(trim($texto));
// Monto a frase apenas com as palavras corretas
$v_frase_correta =
$retorna_resposta->
limpa_texto(trim($semponto),$v_stopwords);

if (count($v_frase_correta)<=0) {
   echo("<html>\n");
   echo("  <head><title>Responde</title></head>\n");
   echo("  <body link=#0000ff vlink=#0000ff bgcolor=#33CCFF
   onLoad=\"window.moveTo(150,80);self.focus(); \">\n");
   echo("  <form name=entrar action=entrada_dados.php
   method=post>\n");
   echo("  <table width=100%><tr><td>\n");
   echo("  <tr><td>\n");
   echo("  <input class=text type=button value='Voltar'
   onclick=self.close();></td></tr></table>\n");
   die("   Não encontrei palavras chaves para pesquisa");
   echo("  </td><td align=right>\n");
   echo("  </td><td align=right>\n");
   echo("  </td><td align=right>\n");
   echo("  </body>\n");
   echo("</html>\n");

} else {
```

```
   // Aqui tenho um vetor já com as
   // palvras corretas $v_frase_correta

   /* O objetivo é verificar quantas palvras existem
   de cada palavra passada
      na pergunta e ordenar o vetor de resposta */

   $v_quantidade_palavras =
   $retorna_resposta->calcula_mais_palavras($v_frase_correta);

   /* Verifica se pelo menos 1 palavra tem pelo menos 1 ocorrencia */
   $tem            =
   $retorna_resposta->verifica_se_tem($v_quantidade_palavras);

   if ($tem) {
      $rs   =
      $retorna_resposta->retorna_docs($v_quantidade_palavras);
   }

   // Vou mostrar
   echo("<html>\n");
   echo("  <head><title>Responde</title></head>\n");

   echo("  <body link=#0000ff vlink=#0000ff bgcolor=#33CCFF
   onLoad=\"window.moveTo(150,80);self.focus(); \">\n");
   //echo("  <body link=#0000ff vlink=#0000ff
   bgcolor=#33CCFF>\n");
   echo("  <form name=entrar action=entrada_dados.php
   method=post>\n");
   echo("     <table width='100%' border='0'
   bgcolor='#33CCFF' style='text-align: justify'>");
   echo("       <tr><td>");
   echo("         <b><FONT SIZE='+1' COLOR='#000000'
   align='justify'> Pergunta : ".$texto."</FONT> </b>");
   echo("       </td></tr>");
   echo("     </table>");
   echo("     <HR>");
   /* Tabela para Imprimir as respostas encontradas */
   echo("     <table width='100%' border='1' bgcolor='#33CCFF'
   style='text-align: justify'>");
   if(mysql_num_rows($rs) == 0){
      echo("<tr><td bordercolor='#000000' border='1'>");
      echo("   <b><font size='+2'>Desculpe,
      não tenho sua resposta
      </font></b>");
      echo("</td></tr> \n");
   }
   else {
      $i =1;
      while ($linha = mysql_fetch_array($rs,MYSQL_ASSOC)){

         echo("<tr><td bordercolor='#000000' border='1'>");
//         echo("   <b><font size='+1'>Resposta $i :
".$linha['titulo']."</font></b><br><br>".
```

```
$linha['texto']."<br>");
         echo("   <b><font size='+1'>Resposta $i : ".
         $linha['titulo']."</font></b><br>");
 //        echo("<b><font size='+1'>
 <textarea name='titulo'
 id='titulo' cols='70' rows='2'>Resposta $i : ".
 $linha['titulo'].
 "</textarea> </font></b>");
 //        echo("<br><br>");
         echo("   <textarea name='texto' id='texto'
         cols='90' rows='7'> ".$linha['texto']."
         </textarea>");

         echo("   <HR>");
         echo("</td></tr>");
         $i++;
      }//end while
   }
   echo("     </table>");
   echo("     <div align='center'> ");
   echo("        <input class=text type=button value='Sair'
   onclick=self.close();></td></tr></table>\n");
   echo("     </div> ");
   echo("  </body>");
   echo("</html>");
}

?>
```

A.7 Programa para fazer a pergunta em PHP

```
Programa que é inserido no TELEDUC no programa da
sala de bate-papo, permitindo que seja feita a pergunta
<?php
  /* Incluido para a Dissertacao de Mestrado de
  Eros Moura em 01/09/05 */

  echo("<script language=\"javascript\"
  type=\"text/javascript\"> \n");
  echo("function Envia_PopUp(Consulta){ \n");
  echo("    var test = document.Consulta.pergunta.value; \n");
  echo("    url=window.open('retorna_resposta.php?
  pergunta='+test,'Resposta',
  'width=800, height=600,scrollbars=yes, resizable=yes '); \n");
  echo("  return true; \n");
  echo("} \n");
  echo("</script> \n");

  echo("<form name='Consulta'> \n");
  echo("   Entre com a pergunta  <input type='text'
  size=50 name='pergunta' >\n");
```

```
  echo("     <input name='md' type='button'
  value='Fazer Pergunta a Base de Conhecimento'
  onClick=\"Envia_PopUp(Consulta);\" >" );
  echo("</form> \n");

  /*-----------------------------------------------
  ---------------------*/

?>
```

ANEXO B – Programas que utilizam MD com mecanismo para acesso aos dados

Relação dos programas que implementam mineração de dados como mecanismo de acesso aos dados. Estes programas foram desenvolvidos na linguagem Java, utilizando como banco de dados o Mysql.

B.1 Programa principal da ferramenta feito em Java

```
/*
 * DMServ.java
 *
 * Created on 28 de Agosto de 2005, 00:46
 */

package dm;

/**
 *
 * @author  erosmoura
 */
public class DM extends javax.swing.JFrame {

    /** Creates new form DMServ */
    public DM() {
        initComponents();
    }

    /** This method is called from within
    the constructor to
     * initialize the form.
     * WARNING: Do NOT modify this code.
     The content of this method is
     * always regenerated by the Form Editor.
     */
    // <editor-fold defaultstate="collapsed" desc="
```

```
Generated Code ">
//GEN-BEGIN:initComponents
private void initComponents() {
    jFrame1 = new javax.swing.JFrame();
    jMenuBar1 = new javax.swing.JMenuBar();
    jMenu1 = new javax.swing.JMenu();
    jMenuItem1 = new javax.swing.JMenuItem();
    jMenu2 = new javax.swing.JMenu();
    jMenuItem2 = new javax.swing.JMenuItem();
    jMenu3 = new javax.swing.JMenu();
    jMenuItem3 = new javax.swing.JMenuItem();
    jMenu4 = new javax.swing.JMenu();
    jMenuItem4 = new javax.swing.JMenuItem();
    jMenu5 = new javax.swing.JMenu();

    getContentPane().setLayout(
    new org.netbeans.lib.awtextra.
    AbsoluteLayout());

    setDefaultCloseOperation(
    javax.swing.WindowConstants.EXIT_ON_CLOSE);
    setTitle("DM Serv");
    setAlwaysOnTop(true);
    setName("principal");
    getAccessibleContext().
    setAccessibleDescription("DM Serv - Principal");
    jMenu1.setText("Arquivo");
    jMenuItem1.setText("Sair");
    jMenuItem1.setActionCommand("Item");
    jMenuItem1.addActionListener(
    new java.awt.event.ActionListener() {
        public void actionPerformed(
        java.awt.event.ActionEvent evt) {
            jMenuItem1ActionPerformed(evt);
        }
    });
    jMenuItem1.addMenuKeyListener(
    new javax.swing.event.MenuKeyListener() {
        public void menuKeyPressed(
        javax.swing.event.MenuKeyEvent evt) {
            jMenuItem1MenuKeyPressed(evt);
        }
        public void menuKeyReleased(
        javax.swing.event.MenuKeyEvent evt) {
        }
        public void menuKeyTyped(
        javax.swing.event.MenuKeyEvent evt) {
        }
    });
    jMenuItem1.addMouseListener(
    new java.awt.event.MouseAdapter() {
        public void mouseClicked(
        java.awt.event.MouseEvent evt) {
            jMenuItem1MouseClicked(evt);
```

```
    }
});

jMenu1.add(jMenuItem1);

jMenuBar1.add(jMenu1);

jMenu2.setText("Acesso a Dados");
jMenuItem2.setText("Busca os Dados");
jMenuItem2.addActionListener(
new java.awt.event.ActionListener() {
    public void actionPerformed(
    java.awt.event.ActionEvent evt) {
        jMenuItem2ActionPerformed(evt);
    }
});

jMenu2.add(jMenuItem2);

jMenuBar1.add(jMenu2);

jMenu3.setText("Minera Dados");
jMenu3.addActionListener(
new java.awt.event.ActionListener() {
    public void actionPerformed(
    java.awt.event.ActionEvent evt) {
        jMenu3ActionPerformed(evt);
    }
});
jMenu3.addMouseListener(
new java.awt.event.MouseAdapter() {
    public void mouseClicked(
    java.awt.event.MouseEvent evt) {
        jMenu3MouseClicked(evt);
    }
});

jMenuItem3.setText("Processa Mining");
jMenuItem3.addActionListener(
new java.awt.event.ActionListener() {
    public void actionPerformed(
    java.awt.event.ActionEvent evt) {
        jMenuItem3ActionPerformed(evt);
    }
});

jMenu3.add(jMenuItem3);

jMenuBar1.add(jMenu3);

jMenu4.setText("Manuten\u00e7\u00e3o na Interface");
jMenuItem4.
setText("Manuten\u00e7\u00e3o em Respostas");
jMenuItem4.addActionListener(
```

```
        new java.awt.event.ActionListener() {
            public void actionPerformed(
            java.awt.event.ActionEvent evt) {
                jMenuItem4ActionPerformed(evt);
            }
        });

        jMenu4.add(jMenuItem4);

        jMenuBar1.add(jMenu4);

        jMenu5.setText("Ajuda");
        jMenuBar1.add(jMenu5);

        setJMenuBar(jMenuBar1);

        java.awt.Dimension screenSize =
        java.awt.Toolkit.
        getDefaultToolkit().getScreenSize();
        setBounds((screenSize.width-631)/2,
        (screenSize.height-457)/2, 631, 457);
    }
    // </editor-fold>//GEN-END:initComponents

    private void jMenuItem4ActionPerformed(
    java.awt.event.ActionEvent evt)
    {//GEN-FIRST:event_jMenuItem4ActionPerformed
    // TODO add your handling code here:
       Respostas respostas = new Respostas();
       respostas.setVisible(true);
    }//GEN-LAST:event_jMenuItem4ActionPerformed

    private void jMenu3MouseClicked(
    java.awt.event.MouseEvent evt)
    {//GEN-FIRST:event_jMenu3MouseClicked
       // TODO add your handling code here:

    }//GEN-LAST:event_jMenu3MouseClicked

    private void jMenuItem3ActionPerformed(
    java.awt.event.ActionEvent evt)
    {//GEN-FIRST:event_jMenuItem3ActionPerformed
       // TODO add your handling code here:
         Processa processa = new Processa();
         processa.setVisible(true);
    }//GEN-LAST:event_jMenuItem3ActionPerformed

    private void jMenu3ActionPerformed(
    java.awt.event.ActionEvent evt)
    {//GEN-FIRST:event_jMenu3ActionPerformed
       // TODO add your handling code here:

    }//GEN-LAST:event_jMenu3ActionPerformed
```

```
    private void jMenuItem2ActionPerformed(
    java.awt.event.ActionEvent evt)
    {//GEN-FIRST:event_jMenuItem2ActionPerformed
// TODO add your handling code here:
        //Gera_Arquivo_WEKA arquivo = new Gera_Arquivo_WEKA();
        Gera_Dados arquivo = new Gera_Dados();
        arquivo.setVisible(true);
    }//GEN-LAST:event_jMenuItem2ActionPerformed

    private void jMenuItem1MouseClicked(
    java.awt.event.MouseEvent evt)
    {//GEN-FIRST:event_jMenuItem1MouseClicked
// TODO add your handling code here:

    }//GEN-LAST:event_jMenuItem1MouseClicked

    private void jMenuItem1MenuKeyPressed(
    javax.swing.event.MenuKeyEvent evt)
    {//GEN-FIRST:event_jMenuItem1MenuKeyPressed
// TODO add your handling code here:

    }//GEN-LAST:event_jMenuItem1MenuKeyPressed

    private void jMenuItem1ActionPerformed(
    java.awt.event.ActionEvent evt)
    {//GEN-FIRST:event_jMenuItem1ActionPerformed
// TODO add your handling code here:
    System.exit(0);

    }//GEN-LAST:event_jMenuItem1ActionPerformed

    /**
     * @param args the command line arguments
     */
    public static void main(String args[]) {
        java.awt.EventQueue.invokeLater(new Runnable() {
            public void run() {
                new DM().setVisible(true);
            }
        });
    }

    // Variables declaration -
    do not modify//GEN-BEGIN:variables
    private javax.swing.JFrame jFrame1;
    private javax.swing.JMenu jMenu1;
    private javax.swing.JMenu jMenu2;
    private javax.swing.JMenu jMenu3;
    private javax.swing.JMenu jMenu4;
    private javax.swing.JMenu jMenu5;
    private javax.swing.JMenuBar jMenuBar1;
    private javax.swing.JMenuItem jMenuItem1;
    private javax.swing.JMenuItem jMenuItem2;
    private javax.swing.JMenuItem jMenuItem3;
```

```
    private javax.swing.JMenuItem jMenuItem4;
    // End of variables declaration//GEN-END:variables

}
```

B.2 Programa gera arquivo WEKA feito em Java

```
Programa que lê uma base de dados e gera arquivo
no formato do WEKA feito em Java
/*
 * Gera_Dados.java
 *
 * Created on 28 de Agosto de 2005, 02:35
 */

package dm;

import java.io.*;
import java.util.*;
import weka.core.*;
import java.sql.*;
import weka.experiment.DatabaseUtils;
import weka.experiment.InstanceQuery;

/**
 *
 * @author  erosmoura
 */
public class Gera_Dados extends javax.swing.JDialog {

    /** Creates new form Gera_Dados */
    public Gera_Dados() {
        initComponents();
        gera_arquivo_WEKA();
    }

    /** This method is called from within the constructor to
     * initialize the form.
     * WARNING: Do NOT modify this code. The content of this method is
     * always regenerated by the Form Editor.
     */
   private void initComponents()
      {//GEN-BEGIN:initComponents
      jFrame1 = new javax.swing.JFrame();
      jLabel1 = new javax.swing.JLabel();
      jLabel2 = new javax.swing.JLabel();
      jButton1 = new javax.swing.JButton();

      getContentPane().setLayout(new org.netbeans.lib.awtextra.
      AbsoluteLayout());
```

```
      setTitle("Gera Dados lendo de uma base de Dados");
      setAlwaysOnTop(true);
      jLabel1.setText("Gerando :  ");
      getContentPane().add(jLabel1, new org.netbeans.lib.awtextra.
      AbsoluteConstraints(40, 90, -1, -1));

      jLabel2.setHorizontalAlignment(javax.swing.SwingConstants.LEFT);
      jLabel2.setText("....");
      getContentPane().add(jLabel2, new org.netbeans.lib.awtextra.
      AbsoluteConstraints(100, 90, 330, -1));

      jButton1.setText("Fechar");
      jButton1.addActionListener(new java.awt.event.ActionListener() {
         public void actionPerformed(java.awt.event.ActionEvent evt) {
            jButton1ActionPerformed(evt);
         }
      });

      getContentPane().add(jButton1, new org.netbeans.lib.awtextra.
      AbsoluteConstraints(170, 160, -1, -1));

      java.awt.Dimension screenSize =
      java.awt.Toolkit.getDefaultToolkit().getScreenSize();
      setBounds((screenSize.width-487)/2,
      (screenSize.height-329)/2, 487, 329);
   }//GEN-END:initComponents

    private void jButton1ActionPerformed(
    java.awt.event.ActionEvent evt)
    {//GEN-FIRST:event_jButton1ActionPerformed
// TODO add your handling code here:
       this.dispose();
    }//GEN-LAST:event_jButton1ActionPerformed

    /**
     * @param args the command line arguments
     */
/*
    public static void main(String args[]) {
        java.awt.EventQueue.invokeLater(new Runnable() {
            public void run() {
                new Gera_Dados().setVisible(true);
            }
        });
    }
*/
   // Variables declaration - do not modify//GEN-BEGIN:variables
   private javax.swing.JButton jButton1;
   private javax.swing.JFrame jFrame1;
   private javax.swing.JLabel jLabel1;
   private javax.swing.JLabel jLabel2;
   // End of variables declaration//GEN-END:variables

    public void gera_arquivo_WEKA() {
```

```
    try {

      DatabaseUtils conexao;
      InstanceQuery iq;
      Instances instances;
      jLabel2.setText("Criando o arquivo... ");

      PrintStream saida=new PrintStream("exp_dados.arff");

      conexao = new DatabaseUtils();
      conexao.connectToDatabase();

      iq = new InstanceQuery();

      instances = new Instances(iq.retrieveInstances
      ("select palavra_chave1,palavra_chave2,palavra_chave3,
      tem_resposta from respostas"));

      jLabel2.setText("Gerando as Instâncias... ");

      instances.setClassIndex(instances.numAttributes() - 1);
      if (instances.numInstances()>0) {
         jLabel2.setText("Arquivo gerado corretamente : "+
         instances.numInstances()+" Geradas");
      }
      else {
         jLabel2.setText("Não foram geradas Instâncias !");

      }
      System.setOut(saida);
      // Imprime os dados
      //jLabel2.setText("Arquivo gerado corretamente : "+
      instances.numInstances()+" Geradas");
      System.out.println(instances);
      saida.close();


    } catch (Exception e) {
        System.err.println(e.getMessage());
    }


    }

}
```

B.3 Programa que processa a MD feito em Java

```
Programa que processa a mineração de dados feito em Java
/*
```

```
 * Processa.java
 *
 * Created on 11 de Setembro de 2005, 12:32
 */

package dm;

import java.io.*;
import java.util.*;
import weka.core.*;
import weka.experiment.DatabaseUtils;
import weka.experiment.InstanceQuery;
import weka.classifiers.trees.*;
import weka.classifiers.bayes.NaiveBayes;
import weka.classifiers.trees.j48.*;
import weka.clusterers.Clusterer.*;
import weka.clusterers.Cobweb;
import weka.classifiers.trees.ADTree;
import weka.gui.graphvisualizer.GraphVisualizer;
import weka.gui.explorer.ClassifierPanel;
import javax.swing.*;
import java.awt.*;

/**
 *
 * @author  erosmoura
 */
public class Processa extends javax.swing.JDialog {

   /** Creates new form Processa */
   public Processa() {
      initComponents();
      jComboBox1.addItem(makeObj("Árvore-ID3"));
      jComboBox1.addItem(makeObj("Árvore-J48"));
      jComboBox1.addItem(makeObj("Árvore-NBTree"));
      jComboBox1.addItem(makeObj("Bayes-NaiveBayes"));
      jComboBox1.addItem(makeObj("Árvore-RandomTree"));
      jComboBox1.addItem(makeObj("Clusterer-Cobweb"));
      jComboBox1.addItem(makeObj("Árvore-ADTree"));
   }

   private Object makeObj(final String item)
    {
       return new Object() {
       public String toString() { return item; } };
   }

   /** This method is called from within the constructor to
    * initialize the form.
    * WARNING: Do NOT modify this code.
    The content of this method is
    * always regenerated by the Form Editor.
    */
   private void initComponents() {//GEN-BEGIN:initComponents
```

```
jLabel1 = new javax.swing.JLabel();
jButton1 = new javax.swing.JButton();
jLabel2 = new javax.swing.JLabel();
jComboBox1 = new javax.swing.JComboBox();
jPanel1 = new javax.swing.JPanel();
jScrollPane1 = new javax.swing.JScrollPane();
texto = new javax.swing.JTextPane();
jButton2 = new javax.swing.JButton();

getContentPane().setLayout(new org.netbeans.lib.awtextra.
AbsoluteLayout());

setDefaultCloseOperation(javax.swing.WindowConstants.
DISPOSE_ON_CLOSE);
setTitle("Processa Mining");
setAlwaysOnTop(true);
jLabel1.setText("Processa o Mining ....");
getContentPane().add(jLabel1, new org.netbeans.lib.awtextra.
AbsoluteConstraints(270, 10, -1, -1));

jButton1.setText("Processar");
jButton1.addMouseListener(new java.awt.event.MouseAdapter() {
   public void mouseClicked(java.awt.event.MouseEvent evt) {
      jButton1MouseClicked(evt);
   }
});

getContentPane().add(jButton1, new org.netbeans.lib.awtextra.
AbsoluteConstraints(200, 330, -1, -1));

jLabel2.setHorizontalAlignment(javax.swing.SwingConstants.CENTER);
jLabel2.setText("...");
getContentPane().add(jLabel2, new org.netbeans.lib.awtextra.
AbsoluteConstraints(150, 90, 270, -1));

jComboBox1.setEditor(jComboBox1.getEditor());
jComboBox1.addActionListener(new java.awt.event.ActionListener() {
   public void actionPerformed(java.awt.event.ActionEvent evt) {
      jComboBox1ActionPerformed(evt);
   }
});

getContentPane().add(jComboBox1, new org.netbeans.lib.awtextra.
AbsoluteConstraints(230, 50, 180, -1));

jPanel1.setLayout(new org.netbeans.lib.awtextra.AbsoluteLayout());

texto.setText("
                  ");
jScrollPane1.setViewportView(texto);

jPanel1.add(jScrollPane1, new org.netbeans.lib.awtextra.
AbsoluteConstraints(20, 20, 590, 160));
```

```
   getContentPane().add(jPanel1, new org.netbeans.lib.awtextra.
   AbsoluteConstraints(10, 110, 620, 200));

   jButton2.setText("Fechar");
   jButton2.addMouseListener(new java.awt.event.MouseAdapter() {
      public void mouseClicked(java.awt.event.MouseEvent evt) {
         jButton2MouseClicked(evt);
      }
   });

   getContentPane().add(jButton2, new org.netbeans.lib.awtextra.
   AbsoluteConstraints(350, 330, -1, -1));

   java.awt.Dimension screenSize =
   java.awt.Toolkit.getDefaultToolkit().
   getScreenSize();
   setBounds((screenSize.width-649)/2,
   (screenSize.height-413)/2, 649, 413);
}//GEN-END:initComponents

private void jButton2MouseClicked(java.awt.event.MouseEvent evt)
{//GEN-FIRST:event_jButton2MouseClicked
   // TODO add your handling code here:
     this.dispose();
}//GEN-LAST:event_jButton2MouseClicked

 private void jComboBox1ActionPerformed(
 java.awt.event.ActionEvent evt) {
 //GEN-FIRST:event_jComboBox1ActionPerformed
     // TODO add your handling code here:
 }//GEN-LAST:event_jComboBox1ActionPerformed

private void jButton1MouseClicked(
java.awt.event.MouseEvent evt)
{//GEN-FIRST:event_jButton1MouseClicked
   // Read all the instances in the file
   try {
      FileReader reader = new FileReader("exp_dados.arff");

      try {
         Instances instances = new Instances(reader);
         // Make the last attribute be the class
         instances.setClassIndex(instances.numAttributes() - 1);

         // Print header and instances.
         jLabel2.setText("Quantidade de Instâncias geradas : "+
         instances.numInstances());

         switch (jComboBox1.getSelectedIndex()) {
            case 0: {
               Id3 dados = new Id3();
               try {
                  dados.buildClassifier(instances);
                  texto.setText(dados.toString());
```

```
   }
   catch (Exception e) {
      jLabel2.setText("Não foi possível
      fazer a Classificação");
   }
   // double[] x = dados.classifyInstance();
   break;
}
case 1: {
   J48 dados = new J48();
   try {
      dados.buildClassifier(instances);
      texto.setText(dados.toString()+dados.graph());
   }
   catch (Exception e) {
      jLabel2.setText("Não foi possível
      fazer a Classificação");
   }
   // double[] x = dados.classifyInstance();
   break;
}
case 2: {
   NBTree dados = new NBTree();
   try {
      dados.buildClassifier(instances);
      texto.setText(dados.toString());
   }
   catch (Exception e) {
      jLabel2.setText("Não foi possível
      fazer a Classificação");
   }
   // double[] x = dados.classifyInstance();
   break;
}
case 3: {
   NaiveBayes dados = new NaiveBayes();
   try {
      dados.buildClassifier(instances);
      texto.setText(dados.toString());
   }
   catch (Exception e) {
      jLabel2.setText("Não foi possível
      fazer a Classificação");
   }
   break;
}
case 4: {
   RandomTree dados = new RandomTree();
   try {
      dados.buildClassifier(instances);
      texto.setText(dados.toString());
   }
   catch (Exception e) {
      jLabel2.setText("Não foi possível
```

```
                    fazer a Classificação");
                }
                // double[] x = dados.classifyInstance();
                break;
            }
            case 5: {
                Cobweb dados = new Cobweb();
                try {
                    dados.buildClusterer(instances);
                    //dados.graph();
                    texto.setText(dados.toString());
                }
                catch (Exception e) {
                    jLabel2.setText("Não foi possível
                    fazer a Classificação");
                }
                // double[] x = dados.classifyInstance();
                break;
            }
            case 6: {
                ADTree dados = new ADTree();
                try {
                    dados.buildClassifier(instances);
                    //dados.graph();
                    //texto.setText(dados.toString());
                    //GraphVisualizer grafico = new GraphVisualizer();
                    javax.swing.JFrame jf = new javax.swing.JFrame
                    ("Weka Explorer: Classifier");
                    jf.getContentPane().setLayout(new BorderLayout());
                    ClassifierPanel sp = new ClassifierPanel();
                    jf.getContentPane().add(sp, BorderLayout.CENTER);
                    weka.gui.LogPanel lp = new weka.gui.LogPanel();
                    sp.setLog(lp);
                    jf.getContentPane().add(lp, BorderLayout.SOUTH);
                    jf.pack();
                    jf.setSize(800, 600);
                    jf.setVisible(true);


              sp.setInstances(instances);

                    String teste = "Teste";
//                    sp.visualizeTree(dados.toString(),teste);

//visualizeTree
//protected void
//visualizeTree(String dottyString,
//   String treeName)Pops up a TreeVisualizer
for the classifier from the currently selected
item in the results list

//Parameters:
//dottyString - the description of the tree in dotty format
//treeName - the title to assign to the display
```

```
                    texto.setText(dados.graph());
                }
                catch (Exception e) {
                    jLabel2.setText("Não foi possível fazer a
                    Classificação");
                }
                // double[] x = dados.classifyInstance();
                break;
            }
            default: {
                jLabel2.setText("Sem Opção");
            }
         }

      }
      catch (IOException e) {
         System.out.println("\nDataset:\n");
      }
   }
   catch (FileNotFoundException e) {
      System.out.println("\nDataset:\n");
   }

}//GEN-LAST:event_jButton1MouseClicked

/**
 * @param args the command line arguments
 */
public static void main(String args[]) {
   java.awt.EventQueue.invokeLater(new Runnable() {
      public void run() {
         new Processa().setVisible(true);
      }
   });
}



// Variables declaration - do not modify//GEN-BEGIN:variables
private javax.swing.JButton jButton1;
private javax.swing.JButton jButton2;
private javax.swing.JComboBox jComboBox1;
private javax.swing.JLabel jLabel1;
private javax.swing.JLabel jLabel2;
private javax.swing.JPanel jPanel1;
private javax.swing.JScrollPane jScrollPane1;
private javax.swing.JTextPane texto;
// End of variables declaration//GEN-END:variables

}
```

B.4 Programa que retorna uma resposta feito em PHP

```
Programa que retorna uma resposta, utilizando a árvore gerada
pela camada de inteligência feito em PHP}
<?

class Retorna_Palavra {
   function conecta() {
   // Conecta a base de Dados
      require_once('conecta.inc');
   }

   function troca_pontos($text){

      //$badword_array = explode(" ",file_get_contents("pontos.txt"));
      $pontos = explode(" ","\, \. \? \;");
      foreach($pontos as $key=>$val)
      {
         $text = eregi_replace(trim($val),' ',$text);
      }
      return $text;
   }

   function tira_palavras($text,$p_palavras_fora){
      $retorno        = '';
      // $palavras_fora  = explode(" ",file_get_contents("fora.txt"));
      $palavras_fora  = explode(" ",$p_palavras_fora);
      //  $array_palavras = str_word_count($text,1);
      $array_palavras = explode(" ", $text);

      foreach($array_palavras as $i=>$palavra)
      {
         // Verifica se a palvra da vez é uma que não serve
         foreach($palavras_fora as $key=>$val)
         {
            // passar para UPPER para comparar
            if (strtoupper(trim($palavra))==strtoupper(trim($val))) {
               // Se não servir coloco "" substituindo
               $nova = "";
               break;
            }
            else {
               // Se não encontrar no vetor palavras_fora é
               porque serve, então continua
               $nova = $palavra." ";
            }
         }
         // Se encontrar uma palvra que serve adiciono a mesma
         na frase correta
         if ($nova <> "") {
            $retorno = $retorno . $nova;
         }
```

```
    }
    return $retorno;
}

function carrega_stopwords(){
    //Executa a query
    $rs = $this->exec_query("Select * from stopwords;");
    $i = 0;
    $v_stopwords[$i]="";
    while ($linha = mysql_fetch_array($rs,MYSQL_ASSOC)){

        $v_stopwords[$i] = $linha['palavra'];
        $i++;
    }
    return($v_stopwords);
}

function pesquisa($pergunta){
    $palavras  = explode(" ",$pergunta);
    // Conta quantas palavras há no vetor de perguntas
    $q_palavras = count($palavras);

    /* Devo definir como fazer a query. A palavra $palavra[0],
    pode ser a mais significante ou não. Como fazer?*/

    /*
    select CONCAT("select titulo, respota from perguntas",
    "where ",coluna1," = '$palavras[0]'",
    " and ",coluna2," = '$pc2' ",
    " and ",coluna3," = '$pc3' "
    " and ",coluna4," = '$pc4' ",
    " and ",coluna5," = '$pc5' ") sql
    from resultado_md
    */
}

function retorna_query2($pc){
    $tamanho = count($pc);
    if ($tamanho == 0) {
        die("Erro-Palavras Chaves para pesquisa igual a 0");
    }
    elseif ($tamanho < 2) {
        $query = "select CONCAT(\"select titulo,resposta from perguntas \",";
        $query = $query . "\"where \",coluna1,\" = '$pc[0]' \", ";
        $query = $query . " ) sql from resultado_md ";
    }
    else {
        $query = "select CONCAT(\"select titulo,resposta
        from perguntas \",";
        $query = $query . "\"where \",coluna1,\" = '$pc[0]' \", ";
        for ($i=1;$i<=$tamanho-2;$i++) {
            $coluna = $i+1;
            $query = $query . " \" and \",coluna$coluna,\" =
            '$pc[$i]' \" ";
```

```
      }
      $query = $query . " ) sql from resultado_md ";
   }
   return($query);
}

// Completa
/* a idéia é pegar o primeiro campo e fazer = valor x
OR valor y OR valor Z, etc */
// $pc - palavra para pesquisa e
$nc - numero de colunas (inicialmente é 5)

/* Tenho uma tabela chamada resultado_md, essa tabela
guarda a ordem correta (calculado pelo Mining)
que devo colocar as colunas no
where. Todas as colunas dessa tabela tem o nome
palavra_chaveX (onde X é um número de 1 a 5).
Então, sabendo qual é a ordem podemos ir retirando
a restrição até achar uma resposta

A ordem, se o mining estiver correto, é do menos
restritivo ao mais restritivo
*/

function retorna_query($pc,$nc,$outros){

   // Vou buscar na tabela resultado_md qual foi a
   ordem das colunas encontradas pelo data mining
   $query = "Select * from resultado_md";
   // Executando a query
   $ordem_colunas = $this->exec_query($query);
   $l_ordem_colunas = mysql_fetch_array($ordem_colunas);


   $tamanho = count($pc);
   if ($tamanho == 0) {
      die("Erro-Palavras Chaves para pesquisa igual a 0");
   }
   else {
      // deve ser igual a 1
      $tamanho2 = count($ordem_colunas);
      if ($tamanho2 == 0) {
         die("Erro-Não foi encontrado o resultado do
         Data Mining na tabela resultado_md");
      }

      /* Fazendo ajuste no tamanho do vetor pois em PHP o
       vetor começa em 0 */
      $tamanho = $tamanho -1;

      $query = "select titulo,resposta from respostas where ";

      // As colunas da tabela resulta_md comecam em 1 e vão até 5
      for ($c=1;$c<=$nc;$c++) {
```

```
            $query = $query . $l_ordem_colunas[$c] . " IN (";

            for ($i=0;$i<=$tamanho;$i++) {
               // Ainda tem mais colunas para serem colocados no IN
               if ($i < $tamanho) {
                  $query = $query . " '$pc[$i]',";
               }
               else {
                  // É a última coluna do IN
                  if ($outros == "N") {
                     $query = $query . " '$pc[$i]' ";
                  }
                  else {
                     // É necessário esse controle pois se não
                     fica no final (na ultima tentativa
                     // where palavra_chaveX in (x,y,z,' '),
                     assim todas as linhas com palavra_chaveX
                     com espaço são aceitas
                     if ($nc > 1) {
                        $query = $query . " '$pc[$i]', ' ' ";
                     }
                     else {
                        $query = $query . " '$pc[$i]' ";
                     }
                  }
               }
            }
            // Fim desse IN com todas as colunas
            $query = $query . " ) ";

            // Verifica se haverá mais AND
            if ($c <= $nc-1) {
               $query = $query . " AND ";
            }
         }
      }
      //echo "A query $query e a quantidade de PR $nc <br> $outros<br>";

      return($query);

   }


   function retorna_query_pesq($query) {
      // Executo a query
      $rs = $this->exec_query($query);
      $linha = mysql_fetch_array($rs,MYSQL_ASSOC);
      $retorno = $linha['sql'];
      if (mysql_num_rows($rs== 0)) {
         die("Erro-Não encontrei o SQL do Data Mining");
      }
      else {
         return($retorno);
      }
```

```
   }

}
$palavra = new Palavra();

$palavra->conecta();
$query        = "Select * from palavras_fora;";
$palavras_fora= $palavra->carrega_palavras_fora($query);

$texto        = $_GET['pergunta'];

// Tiro pontos, virgulas, etc
$semponto     = $palavra->troca_pontos(trim($texto));
// Monto a frase apenas com as palavras corretas
$frase_correta = $palavra->tira_palavras(trim($semponto),$palavras_fora);
if (strlen(trim($frase_correta))==0) {
   echo("<html>\n");
   echo("  <head><title>Responde</title></head>\n");
   echo("  <body link=#0000ff vlink=#0000ff bgcolor=#33CCFF
   onLoad=\"window.moveTo(300,150);self.focus(); \">\n");
   echo("  <form name=entrar action=md.php method=post>\n");
   echo("  <table width=100%><tr><td>\n");
   echo("  <tr><td>\n");
   echo("  <input class=text type=button value='Voltar'
   onclick=self.close();></td></tr></table>\n");
   die("   Não encontrei palavras chaves para pesquisa");
   echo("  </td><td align=right>\n");
   echo("  </td><td align=right>\n");
   echo("  </td><td align=right>\n");
   echo("  </body>\n");
   echo("</html>\n");

}
else {
   // Aqui tenho um vetor já com as palvras corretas
   $pc            = explode(" ", trim($frase_correta));

   $nc = 5;

   /* o metodo retorna_query tem a funcao de montar a query
   utilizando a ordem do DM com as palavras selecionadas.

   Inicialmente, são passados o parametro maximo de 5 campos
   (pois são 5 palavras chaves no máximo.

   Encontrar uma pergunta com as 5 palavras chaves seria o máximo de
   especificação possível.

   Quando isso não é possível, vai-se retirando um campo de pesquisa
   até que se encontre pelo menos 1 resposta

   O parametro OUTROS foi criado para permitir inicialmente uma pesquisa
   com apenas as palavras chaves, em uma
   situação que todas as palavras chaves teriam sido passadas (as 5).
```

```
Quando não é encontrada uma resposta,
chamo novamente a função passando OUTROS=S, dessa forma
será levado em conta as respostas sem todas as palavras
chaves, com a utilização do espaço " "
OBS: para funcionar corretamente as palavras_chaves não
utilizadas na pergunta deve ser colocado um espaço ' '
*/

/* tentando apenas com as palvras chaves */
$nc = 5;     // Numero máximo pois são 5 palavras chaves
do {         // fará ao menos 1 vez,
e apenas 1 vez se encontrar com 5
   $outros="S";
   // Monto a query para montar a query definitiva da pergunta
   $query       = $palavra->retorna_query($pc,$nc,$outros);
   // $pc é o vetor com as palavras da pergunta
   //echo "COM: ".$query."<br><br>";
   // Executo a query para definitiva

   $rs  = $palavra->exec_query($query);

   $linhas = mysql_num_rows($rs);

   $nc--;
} while ((mysql_num_rows($rs) == 0) AND ($nc >= 1));

/* Tentando com os campos com strings fazias */
$nc = 5;     // Numero máximo pois são 5 palavras chaves
if (mysql_num_rows($rs) == 0) {
   do {         // fará ao menos 1 vez,
   e apenas 1 vez se encontrar com 5
      $outros="N";
      // Monto a query para montar a query definitiva da pergunta
      $query       = $palavra->retorna_query($pc,$nc,$outros);
      // $pc é o vetor com as palavras da pergunta
      //echo "SEM: ".$query."<br><br>";

      // Executo a query para definitiva

      $rs  = $palavra->exec_query($query);

      $linhas = mysql_num_rows($rs);

      $nc--;
   } while ((mysql_num_rows($rs) == 0) AND ($nc >= 1));
}


// Vou mostrar
echo("<html>\n");
echo("  <head><title>Responde</title></head>\n");

echo("  <body link=#0000ff vlink=#0000ff bgcolor=#33CCFF
onLoad=\"window.moveTo(300,150);self.focus(); \">\n");
```

```
   echo("  <form name=entrar action=md.php method=post>\n");
   echo("     <table width='100%' border='0' bgcolor='#33CCFF'
   style='text-align: justify'>\n");
   echo("       <tr><td>\n");
   echo("         <b><FONT SIZE='+2' COLOR='#000000' align='justify'>
   Pergunta : ".$texto." </FONT> </b>");
   echo("       </td></tr> \n");
   echo("     </table>");
   echo("     <HR>");
   echo("     <HR>");
   /* Tabela para Imprimir as respostas encontradas */
   echo("     <table width='100%' border='1' bgcolor='#33CCFF'
   style='text-align: justify'>\n");
   if(mysql_num_rows($rs) == 0){
      echo("<tr><td bordercolor='#000000' border='1'>\n");
      echo("   <b><font size='+2'>Desculpe, não tenho sua resposta
      </font></b>");
      echo("</td></tr> \n");
   }
   else {
      $i =1;
      while ($linha = mysql_fetch_array($rs,MYSQL_ASSOC)){

         echo("<tr><td bordercolor='#000000' border='1'>\n");
         echo("   <b><font size='+1'>Resposta $i : </font>
         <font>".$linha['titulo']."</font></b><br><br>".
         $linha['resposta']."<br>");
         echo("   <HR>");
         echo("</td></tr> \n");
         $i++;
      }//end while
   }
   echo("     </table>");
   echo("     <div align='center'> ");
   echo("        <input class=text type=button value='Sair'
   onclick=self.close();></td></tr></table>\n");
   echo("     </div> ");
   echo("  </body>\n");
   echo("</html>\n");
}

?>
```

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo