UMA ANÁLISE DOS PROCEDIMENTOS DE MINERAÇÃO DE TEXTOS NO SGBD ORACLE 10G

Luiz Claudio Marini Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____

Profa. Beatriz de Souza Leite Pires de Lima, D.Sc.

_____

Profa. Myrian Christina de Aragão Costa, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

JUNHO DE 2008

# Livros Grátis

http://www.livrosgratis.com.br

Milhares de livros grátis para download.

SILVA, LUIZ CLAUDIO MARINI

Uma Análise dos procedimentos de Mineração de Textos no SGBD Oracle 10G

[Rio de Janeiro] 2008

XI, 85 p., 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Civil, 2008)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Mineração de Textos
2. Oracle 10G
3. Classificação
4. Clusterização
5. Visualização

I. COPPE/UFRJ II. Título (série)

À minha mãe Maria Helena, por sua força e inspiração.

## AGRADECIMENTOS

Agradeço:

ao meu orientador, Professor Nelson Ebecken, pela confiança, estímulo e paciência a mim dedicados, mas sobretudo pelo exemplo de profissionalismo e dedicação sempre presente neste e nos diversos trabalhos que colabora;

ao meu grande amigo Alexandre Soares, por seu incentivo e apoio, sempre presentes, impedindo que as adversidades do trabalho e da vida desmotivassem a minha caminhada;

aos amigos de trabalho na COPPETEC, especialmente a Raquel e ao Vinícius, e tantos outros que torceram pelo meu sucesso e me apoiaram sempre que precisei;

aos colegas de curso Claudia, Claudio e Marcos, pelo companheirismo, apoio e incentivos sempre presentes ao longo dessa trajetória;

a todos os funcionários da COPPE-PEC pelo eficiente apoio administrativo;

à minha família pela compreensão das longas horas de ausência de um convívio tão prazeroso. Em especial à minha mãe Maria Helena, pelas incontáveis palavras de incentivo e motivação;

ao CNPQ, pelo suporte financeiro que viabilizou a realização desta dissertação.

a tantos mais que compartilharam comigo desse caminho;

e, acima de tudo, a Deus.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

# UMA ANÁLISE DOS PROCEDIMENTOS DE MINERAÇÃO DE TEXTOS NO SGBD ORACLE 10G

Luiz Claudio Marini Silva

Junho/2008

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

A extração de conhecimento em bases de dados textuais se tornou uma prática estratégica na esfera empresarial atualmente, uma vez que mais de 80% dos dados produzidos pelos negócios se encontra no formato textual. Essa prática apóia processos de tomada de decisão nas empresas líderes nesse novo milênio, resultando em importantes vantagens competitivas na atual economia globalizada.

O presente trabalho apresenta um estudo aprofundado do ambiente de mineração de textos do Oracle 10G, um Sistema Gerenciador de Banco de Dados (SGBD) muito utilizado no mercado. As abordagens envolvem basicamente tarefas de Classificação, Clusterização e Visualização dos dados textuais.

Como o Oracle é uma ferramenta que, além de armazenar os dados apresenta as abordagens de tratamento deles, o trabalho na etapa de pré-processamento é largamente minimizado, o que confere bastante agilidade ao processo.

Para validar os resultados obtidos com o Oracle, o programa *Poly Analyst* foi utilizado para gerar resultados para comparação.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AN ANALYSIS OF THE TEXT MINING APPROACHES IN THE DATABASE ORACLE 10G

Luiz Claudio Marini Silva

Junho/2008

Advisors: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

The knowledge extraction in textual data became a strategic practice in the business world nowadays, since more than 80% of data produced by businesses are in the textual format. That practice supports decision making processes in the leaders companies on new millennium, resulting in important competitive advantages in the current global economy.

This work presents a detailed study of the text mining environment Oracle 10G, a Managed System of Database very used at the market. The approaches involve tasks of Classification, Clustering and Presenting of the textual data.

As Oracle is a tool that stores data and it also includes data treatment routines, the effort in the initial processing stage is minimized enormously, what aggregate quite agility to the process.

To validate the results obtained with Oracle, the program Poly Analyst was used to generate results for comparison.

# Índice

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviações e Siglas

| | | |
|---|---|---|
| SGDB | - | Sistema Gerenciador de Banco de Dados |
| *KDD* | - | *Knowledge Discovery Database* |
| *KDT* | - | *Knowledge Discovery in Texts* |
| *OT* | - | *ORACLE TEXT* |
| *PA* | - | *Poly Analyst* |
| EC | - | Estudo(s) de Caso(s) |
| EC1 | - | Estudo de Caso 1 |
| EC2 | - | Estudo de Caso 2 |
| EC3 | - | Estudo de Caso 3 |
| EC4 | - | Estudo de Caso 4 |
| EC5 | - | Estudo de Caso 5 |
| *SVM* | - | Algoritmo *Support Vector Machine* |
| AD | - | Algoritmo baseado em Árvores de Decisão |
| RI | - | Recuperação de Informação |
| EI | - | Extração de Informação |
| *SGML* | - | *Standard Generalized Markup Language* |
| *HTML* | - | *HyperText Markup Language* |
| *XML* | - | *Extensible Markup Language* |
| *URL* | - | *Uniform Resource Locator* |
| *SQL* | - | *Structured Query Language* |
| *PL/SQL* | - | *Procedural Language/Structured Query Language* |
| *CSV* | - | *Comma-Separated Values* |

# Capítulo I

# INTRODUÇÃO

Atualmente, no desenvolvimento dos negócios, é produzida uma grande quantidade de dados, gerados em meios eletrônicos, de acordo com o apresentado em [1], que, normalmente, ficam armazenados em estruturas de Bancos de Dados. Esses dados contêm muitas informações a respeito dos negócios que os originaram. Conforme o apresentado em [2], informação é um conjunto de dados ordenados e com alguma significação.

A partir de uma análise cuidadosa dessas informações, é possível extrair um valioso conhecimento sobre o negócio em questão, e esse conhecimento pode ser o combustível para que as instituições se destaquem no atual cenário globalizado. A figura 1.1 ilustra como se dá a extração de conhecimento a partir dos dados.


Figura 2.1 : Extração de Conhecimento a partir dos dados

O mundo vive hoje o que muitos chamam de “a era da informação” ou “a sociedade do conhecimento”, conforme apresentado em [3]. De uma forma geral, os meios de produção e a tecnologia estão acessíveis a todos e a gestão do conhecimento parece ser o grande diferencial entre as empresas de ponta nessa nova configuração. De acordo com o exposto em [4], no século XXI as empresas líderes em seus segmentos deverão ser aquelas que tiverem condições de criar produtos e serviços intensivos em conhecimento.

Um problema corrente que as organizações enfrentam nos dias de hoje é como tratar esse volume enorme de dados gerados pelos negócios. Torna-se necessário agrupá-los em categorias distintas por assunto ou interesse e, a partir daí, dar tratamento específico a eles.

Uma primeira estratégia seria uma abordagem humana de ordenação e classificação dos documentos. Porém, diante de uma amostra muito grande, isso não parece viável.

Uma outra abordagem seria agrupar automaticamente esses dados através de buscas por palavras-chave nos documentos. Mas essa abordagem também apresenta limitações, principalmente por não levar em consideração o contexto em que os termos se encontram. De acordo com o apresentado em [5], em vários idiomas, um termo ou expressão podem ter significados distintos de acordo com o seu contexto, e, ao não se levar em conta essa situação, comumente incorre-se em erros na associação dos documentos aos grupos.

Conforme apresentado em [6], a análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais apropriadas. Torna-se imprescindível, então, o desenvolvimento de ferramentas que auxiliem o homem, de forma automática e inteligente, na tarefa de analisar, interpretar e relacionar esses dados para que estratégias de ação apropriadas possam ser desenvolvidas, em cada contexto do negócio.

Existe uma área de pesquisa que se desenvolveu muito nos últimos anos, denominada *Knowledge Discovery Database* (*KDD*), que trata da extração de conhecimento em bases de dados estruturados, e que vem despertando grande interesse junto às comunidades científica e industrial. De acordo com a definição apresentada em [7], "*KDD* é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados".

Conforme disposto em [8], mais de 80% dos dados produzidos pelas empresas no desenvolvimento dos negócios se encontra no formato textual. Esses dados são tipicamente não estruturados, sendo semi-estruturados ocasionalmente. Para a extração de conhecimento a partir desse tipo de dado, é necessário todo um tratamento específico na fase de pré-processamento a fim de torná-los estruturados, o que normalmente é uma tarefa que exige um esforço considerável.

Dentro de todo esse contexto ganha força à área de Mineração de Textos (*Text Mining*), que, de acordo com o apresentado em [9], busca extrair padrões e conhecimento interessante e não-trivial de bases de dados textuais ou não estruturadas. Ela tem origem na área de *Data Mining* e *KDD*, sendo também conhecida como *Knowledge Discovery in Texts* (*KDT*).

## 1.1 MOTIVAÇÃO

A maior parte das ferramentas de mineração de textos disponíveis hoje em dia trabalha acoplada a um banco de dados. Esse acoplamento normalmente não é trivial, e requer, na maior parte dos casos, algum tratamento específico nos dados, o que pode ser bastante custoso em termos de trabalho, de acordo com a quantidade de dados a ser tratada.

Uma importante plataforma de banco de dados disponível hoje no mercado é o Oracle. Este Sistema Gerenciador de Banco de Dados (SGBD) é uma ferramenta baseada na arquitetura cliente/servidor, robusta, e largamente utilizada no mercado.

A partir da sua versão 9, o Oracle passou a incorporar abordagens específicas para o tratamento de bases de dados textuais.

A motivação principal para o desenvolvimento desse trabalho é investigar detalhadamente as abordagens de mineração de textos dispostas no Oracle 10G, uma vez que ele já é a fonte dos dados e não requer, por isso, nenhum acoplamento ou tratamento dos dados a serem utilizados, estando estes carregados no sistema. Isso minimiza enormemente ou mesmo elimina o trabalho na etapa de pré-processamento dos dados, onde, tipicamente, é consumido o maior esforço em todo o processo de mineração de textos.

## 1.2 CONTRIBUIÇÕES

Podemos citar como principais contribuições alcançadas com o desenvolvimento deste trabalho:

- Uma análise detalhada das abordagens dispostas no Oracle 10G para o tratamento de dados textuais, referentes às tarefas de Classificação, Clusterização e Visualização dos dados;
- Uma avaliação ampla e abrangente do desempenho da ferramenta nas tarefas acima expostas;

- A validação dos resultados alcançados com o Oracle 10G a partir do confrontamento com os resultados obtidos com a ferramenta *Poly Analyst*, nas tarefas de Classificação e Clusterização.

## 1.3 ESTRUTURA DA TESE

Esse capítulo tem a intenção de contextualizar de uma forma ampla o tema desse trabalho, apresentando uma abordagem geral dos assuntos aqui tratados.

No capítulo 2 será apresentada uma visão geral de Mineração de Textos, com ênfase nas tarefas abordadas nesta dissertação: Classificação, Clusterização e Visualização.

O capítulo 3 trata das ferramentas utilizadas no desenvolvimento desse trabalho: o Oracle 10G, alvo de toda a investigação realizada aqui; o *Toad*, utilizado como uma interface do Oracle; e o *Poly Analyst*, utilizado com o intuito de validar os resultados obtidos com a ferramenta alvo.

No capítulo 4 estão dispostos os Estudos de Casos elaborados, seus detalhes e configurações, a base de dados utilizada, os resultados obtidos e a sua análise.

O capítulo 5 traz as conclusões tiradas dos Estudos de Casos e as sugestões indicadas para futuros trabalhos nesta área de pesquisa.

# Capítulo II

# MINERAÇÃO DE TEXTOS

Neste capítulo serão apresentadas inicialmente algumas considerações gerais acerca do assunto Mineração de Textos. A seguir serão abordadas as três etapas envolvidas: o Pré-Processamento, a Mineração de Dados Textuais e o Pós-Processamento.

A Mineração de Textos pode ser definida como um conjunto de técnicas para recuperar e extrair informação em dados textuais escritos em linguagem natural, a fim de descobrir conhecimento inovador nos textos, conforme apresentado em [10]. Com o auxílio dessas técnicas é possível manipular mais facilmente informações não estruturadas como notícias, textos em *websites*, *blogs* e documentos em geral.

Historicamente, a importância dessa área ganhou impulso a partir da década de 90, com o crescimento do armazenamento digital, da internet e dos mecanismos de busca. Ao mesmo tempo, analistas começaram a perceber a ausência de ferramentas de mineração de dados para lidar com o ambiente de informações não estruturadas.

Parte importante do processo de mineração de textos é a preparação dos documentos, objetivando armazenar um texto não estruturado em uma base de dados estruturada. Essa operação é necessária para que os algoritmos das ferramentas de mineração possam ser aplicados.

A mineração de dados textuais é uma área mais complexa que a mineração de dados tradicional, pois os dados textuais são inerentemente não estruturados ou, no máximo, semi-estruturados, conforme o apresentado em [11].

Por ser uma área multidisciplinar, o desenvolvimento de tecnologias de mineração de textos requer conhecimentos específicos a respeito de áreas distintas e correlatas, tais como Estatística, Lingüística, Ciência Cognitiva, Processamento em Linguagem Natural, Extração e Recuperação de Informação, Aprendizagem de

Máquina, Inteligência Artificial, Mineração de Dados e Visualização da Informação, além de conhecimentos em Ciência da Computação e Engenharia de Software.

De acordo com o apresentado em [1], existem duas maneiras principais de abordagem de dados textuais: a análise estatística, que se baseia na freqüência dos termos nos textos e é independente do idioma; e a análise semântica, que se baseia na função que os termos desempenham nos textos, é dependente do idioma e apresenta uma complexidade maior.

O processo de mineração de textos tem início com a definição dos objetivos que se pretende alcançar, conforme o apresentado em [11]. É preciso que se tenha um delineamento detalhado do problema a ser tratado. Uma vez definido o escopo do problema, a mineração de textos apresenta tipicamente três grandes etapas:

1. Pré-Processamento: envolve as tarefas relacionadas à captação, à organização e ao tratamento dos dados, objetivando prepará-los para o processamento pelos algoritmos da etapa seguinte;
2. Mineração de Dados Textuais: abrange as tarefas de processamento dos dados textuais para a extração de conhecimentos úteis;
3. Pós-Processamento: envolve o tratamento do conhecimento obtido na etapa anterior, com o intuito de viabilizar a avaliação do conhecimento descoberto. Essa etapa nem sempre é necessária.

Alguns autores, como os apresentados em [12], consideram a coleta dos documentos uma etapa a parte, anterior ao Pré-Processamento. Neste trabalho foi considerado que a aquisição dos documentos integra a etapa de Pré-Processamento.

Cada etapa será melhor detalhada nos tópicos seguintes.

## 2.1 PRÉ-PROCESSAMENTO

Essa etapa possui uma relevância fundamental no processo de descoberta de conhecimento. Para a obtenção de resultados confiáveis na mineração de textos, é essencial a utilização de dados depurados, conforme apresentado em [13]. Entretanto, os dados disponíveis em geral possuem baixa qualidade. Valores ausentes e incorretos,

combinações inexistentes e erros de grafia são comuns e podem alterar os resultados a serem alcançados.

As atividades de obtenção e limpeza dos dados normalmente consomem mais da metade do tempo dedicado ao processo como um todo. Porém, o tratamento inicial dos dados confere maior consistência a eles e pode evitar a obtenção de resultados distorcidos.

Essa etapa envolve desde a seleção dos dados, passando pela eliminação das inconsistências e dos termos repetidos e com pouca ou nenhuma capacidade preditiva, até o ajuste da sua formatação para o processamento seguinte. Ela é responsável por transformar os textos em uma representação estruturada adequada para o processo de mineração. No entanto, durante essa transformação, existe a possibilidade de a informação intrínseca ao conteúdo dos textos ser perdida. Um desafio, nesse caso, é obter uma boa representação dos textos, minimizando a perda de informação.

No processo de tratamento dos dados textuais algumas sub-tarefas podem ser executadas. Cabe ressaltar que as diferentes implementações não necessariamente consideram todas elas. Serão apresentadas a seguir as quatro principais sub-tarefas empregadas.

### 2.1.1 CASE FOLDING

É o procedimento de converter todos os caracteres de um documento para um único tipo de letra, maiúsculo ou minúsculo. Ele confere maior agilidade na análise dos dados através do processo de indexação.

### 2.1.2 STOPWORDS

Com o intuito de desconsiderar da base os termos que não constituem informação relevante nem possuem conteúdo semântico significativo, é carregada no sistema uma lista de termos denominada *Stopwords* (também conhecida como *Stoplist*), contendo elementos a serem descartados, tais como preposições, pronomes, artigos e outras classes de palavras.

A eliminação desses elementos tem o objetivo de filtrar as palavras que aparecem com muita freqüência e não constituem conhecimento nos textos, reduzindo

de forma considerável a quantidade de termos indexados. Com a execução dessa sub-tarefa é possível reduzir o número de termos do texto em mais de 40%, o que agiliza o processamento futuro.

### 2.1.3 STEMMING

*Stemming* é uma técnica de redução de termos a um radical comum, a partir da análise das características gramaticais dos elementos, como grau, número, gênero e desinência. Tem o objetivo de retirar os sufixos e prefixos das palavras, e encontrar a sua forma primitiva. Assim, palavras no plural ou derivadas são reduzidas a um radical único, a sua raiz, simplificando a representação dos termos envolvidos no documento. Isso implica numa única entrada nos índices, aumentando o desempenho do processo.

Dois erros típicos que costumam ocorrer durante o processo são o *Overstemming* e o *Understemming*. O *Overstemming* se dá quando a cadeia de caracteres removida não é um sufixo, mas parte do raiz do termo. Já o *Understemming* ocorre quando um sufixo não é removido completamente. Um desafio corrente aqui é configurar os parâmetros dos algoritmos que executam essa tarefa a fim de que essas distorções sejam evitadas.

### 2.1.4 USO DE UM *THESAURUS*

Um *Thesaurus* pode ser definido como um dicionário controlado que representa hierarquias, abreviações, sinônimos, acrônimos, ortografias alternativas e relacionamentos associativos entre termos, com o intuito maior de apoiar os usuários na recuperação das informações requeridas, conforme apresentado em [10].

A importância do uso de um dicionário desse tipo fica aparente ao se tratar de questões relacionadas à consulta e a indexação em linguagem natural. Diferentes usuários costumam definir a mesma consulta através de termos distintos.

De forma similar ao que ocorre na abordagem da técnica de *Stemming*, no *Thesaurus* vários termos são mapeados para um termo conceito único, que expressa à idéia geral dos elementos. Isso otimiza a tarefa de indexação dos elementos, agregando mais consistência a ela.

O conhecimento sobre o domínio dos documentos é fundamental para a elaboração de um Thesaurus.

## 2.2 MINERAÇÃO DE DADOS TEXTUAIS

Essa é a principal etapa do processo de mineração de textos. Nela ocorre a busca efetiva por conhecimentos novos e úteis a partir dos dados textuais, através da aplicação dos algoritmos, fundamentados em técnicas que procuram, segundo determinados paradigmas, explorar os dados de forma a produzir modelos de conhecimento.

É preciso que se tenha uma idéia clara das tarefas que se pretende realizar, uma vez que cada tipo de tarefa requer um tratamento próprio e extrai padrões diferentes de informação dos textos.

As tarefas de mineração de textos podem ser implementadas a partir de diferentes técnicas, conforme apresentado em [6], [14], [15] e [16]. Essas técnicas podem ser utilizadas isoladamente ou combinadas, gerando sistemas híbridos. Eis algumas delas:

- Métodos Estatísticos;
- Redes Bayesianas;
- Árvores de Decisão;
- Redes Neurais;
- Algoritmos Genéticos;
- Sistemas *Fuzzy*;
- Algoritmo *Support Vector Machine* (*SVM*).

Cada ferramenta de mineração de textos normalmente apresenta uma abordagem própria da implementação dessas técnicas, de acordo com as tarefas a serem realizadas.

Serão abordadas aqui as principais tarefas de mineração de textos desenvolvidas.

### 2.2.1 INDEXAÇÃO

Essa tarefa possibilita uma busca mais eficiente dos registros que correspondem ao que foi especificado na consulta, sem a necessidade de se examinar os documentos inteiros. Ela é equivalente a indexação realizada em dados estruturados em bancos de dados convencionais, e permite que se evite percorrer tabelas inteiras para a adequada recuperação dos registros.

Diferentes tipos de índices podem ser aplicados aos dados. As tarefas de extração de informação a serem realizadas determinam o tipo de índice a ser utilizado.

### 2.2.2 RECUPERAÇÃO DE INFORMAÇÃO

A Recuperação de Informação (RI) é uma tarefa que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. Ela consiste em identificar no conjunto de documentos pesquisado aqueles que atendem às necessidades de informação do usuário.

Os processos de RI devem representar o conteúdo dos documentos e apresentá-los de forma a permitir uma rápida seleção dos itens que satisfaçam a necessidade de informação especificada na consulta.

Essa tarefa pode ser considerada como um passo inicial no processo de mineração de textos.

### 2.2.3 EXTRAÇÃO DE CARACTERÍSTICAS

As tarefas de Extração de Características procuram recuperar dos textos categorias de termos pré-definidas, como nomes de pessoas, lugares, organizações, datas, etc.

Para a identificação dos termos, os algoritmos de Extração de Características podem usar dicionários ou padrões linguísticos pré-definidos. O nome de um lugar, por exemplo, pode não constar no dicionário, e ainda assim o algoritmo ser capaz de determinar que esse elemento seja um nome e, provavelmente, um termo significativo no documento.

Na implementação dessa tarefa podem ser utilizados algoritmos de reconhecimento de padrões em um conjunto de objetos. Também podem ser empregadas diferentes métricas definidoras da importância dos termos.

### 2.2.4 EXTRAÇÃO DE INFORMAÇÃO

A tarefa de Extração de Informação (EI) procura identificar trechos dos documentos que preencham corretamente os campos de um formulário de resultado pré-definido, que determina o tipo de dado a ser extraído.

As abordagens comumente utilizadas na construção de sistemas de EI incluem o uso do processamento de linguagem natural, adequado para tratar textos livres (não estruturados), a engenharia do conhecimento e a aprendizagem automática, mais adequada para o tratamento de textos estruturados ou semi-estruturados.

### 2.2.5 SUMARIZAÇÃO

De forma resumida, conforme apresentado em [17], a Sumarização pode ser definida como um processo de redução da quantidade de texto em um documento, porém mantendo o significado e a coerência originais.

Ela é uma técnica que procura identificar os termos e as frases mais relevantes no documento e produzir um resumo com eles, que permita uma apresentação geral do documento e possibilite uma rápida identificação do assunto abordado.

Diferentes abordagens podem ser adotadas, como a Sumarização por Abstração, que procura reproduzir a sumarização feita por humanos, e a Sumarização por Extração, onde sentenças inteiras são extraídas do documento original, com o intuito de se construir um texto menor, mas que contenha as idéias principais do documento original.

### 2.2.6 CLASSIFICAÇÃO

A Classificação pode ser descrita como um processo de identificação dos principais tópicos de um documento e a sua posterior associação automática a uma ou mais categorias pré-definidas, conforme o apresentado em [18].

As categorias podem ser criadas segundo duas abordagens distintas: na primeira, elas são especificadas manualmente pelo usuário em uma tabela ou em um dicionário, onde são definidos os rótulos das classes e o conjunto de termos específico de cada uma; na segunda, deve ser fornecido ao sistema um conjunto de documentos pré-categorizados para ser utilizado no treinamento da ferramenta de categorização, que analisa estatisticamente os documentos a partir de modelos lingüísticos, e produz, automaticamente, as regras que definem as categorias.

A primeira abordagem apresenta a vantagem de possibilitar a visualização e o controle das regras, permitindo, assim, que o conhecimento de um especialista seja incorporado ao sistema. Já a segunda traz a vantagem da geração automática das regras,

o que pode ser crucial quando não se dispõe de um conhecimento aprofundado a respeito do domínio das categorias.

A Classificação de textos é uma tarefa que pode ser aplicada a diferentes áreas, tais como na filtragem de textos, na organização de documentos e na indexação automática para sistemas de recuperação de informação.

Para uma descrição detalhada dos processos de classificação de textos, consultar [14], [15], [16], [12] e [19].

### 2.2.7 CLUSTERIZAÇÃO

A tarefa de Clusterização, também conhecida como Agrupamento ou Classificação Não Supervisionada, é utilizada para particionar os registros de uma base textual em grupos ou *Clusters*, de modo que os elementos de cada *Cluster* compartilhem um conjunto de propriedades comuns, que os distingam dos elementos dos demais, conforme apresentado em [6]. O objetivo deste processo é maximizar a similaridade dentro do *Cluster* e minimizá-la entre os *Clusters*.

Em geral, essa tarefa requer que o usuário determine o número de grupos a ser considerado. Com base nesse valor, os registros são alocados aos grupos de forma que os que forem similares fiquem nos mesmos grupos e os que forem diferentes em grupos distintos. Uma vez gerados os grupos, é possível fazer uma análise dos elementos que compõe cada um deles, identificando as suas características comuns, e, assim, podendo criar um rótulo para cada grupo.

Conforme apresentado em [20], algumas técnicas foram desenvolvidas para apoiar os usuários na definição do número de *Clusters*, envolvendo a geração de índices de validação do número de *Clusters*. Alguns índices bastante difundidos são os conhecidos como *Calinski Harabasz*, PBM e *Xie-Beni*, conforme apresentado [21], [22] e [23].

Existem abordagens que trabalham com paradigmas diferentes: algumas não requerem que os usuários informem o número de grupos existentes; outras são capazes de gerar *Clusters* hierárquicos.

## 2.3 PÓS-PROCESSAMENTO

Essa etapa envolve a apresentação, a análise e a interpretação dos resultados, a fim de validar as descobertas obtidas na etapa anterior. Nela o especialista em mineração de textos e o especialista no domínio da aplicação podem, a partir da avaliação dos resultados alcançados, definir novas alternativas de investigação dos dados.

Existem abordagens que implementam métricas de avaliação dos resultados, baseadas na noção de relevância dos documentos. Outras disponibilizam técnicas de visualização dos resultados, que são abordadas no tópico seguinte.

### 2.3.1 VISUALIZAÇÃO

As ferramentas de visualização dos resultados apóiam a interpretação e a avaliação do conhecimento extraído. Elas podem disponibilizar técnicas de extração de informações específicas sobre os textos, como a recuperação de um termo dentro do seu contexto no documento, a extração de um sumário do documento (técnica também conhecida como Sumarização), a extração do assunto principal do documento, a extração da lista de temas contidos em um documento, ou ainda a extração de uma versão do documento com os termos especificados em destaque, entre outras.

Essas técnicas eventualmente se utilizam de gráficos em duas e em três dimensões para a apresentação dos resultados, a fim de facilitar ou até mesmo viabilizar a sua compreensão e interpretação.

As ferramentas de visualização não necessariamente implementam todas as técnicas de visualização apresentadas.

# Capítulo III

# FERRAMENTAS UTILIZADAS

Este capítulo apresenta as ferramentas utilizadas neste trabalho: o SGBD Oracle 10G, cujas tarefas de mineração de textos foram o motivo principal da análise realizada neste trabalho; o *Toad*, um programa de gerenciamento de banco de dados, utilizado aqui como uma interface do Oracle; e o *Poly Analyst*, um programa de mineração de textos, utilizado com o intuito de comparar e validar os resultados obtidos com o Oracle.

## 3.1 ORACLE 10G

O Oracle é uma ferramenta de banco de dados baseada na arquitetura cliente/servidor, que utiliza as linguagens *SQL* e *PL/SQL* para a manipulação dos dados.

A figura 3.1 ilustra a interface do *SQL Plus* do Oracle, através do qual são passados os comandos para a execução das instruções pelo banco de dados.

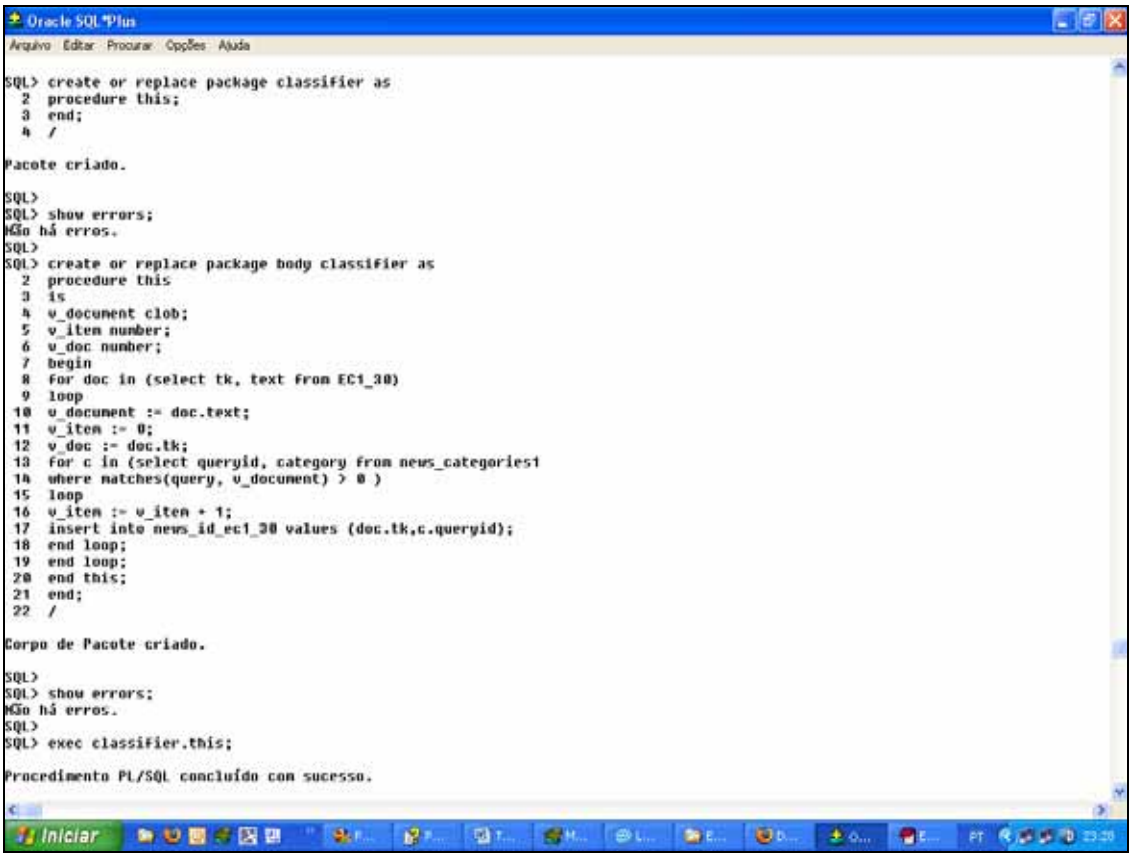


Figura 3.1 : A interface do SQL Plus do Oracle

Um conjunto de rotinas denominado *ORACLE TEXT* (*OT*) foi disponibilizado e incorporado a ferramenta Oracle a partir da versão 9G. Estas rotinas permitem que os usuários realizem consultas e manipulações com dados textuais. O *OT* possibilita, entre outras coisas, pesquisas por palavras ou temas contidos nos textos, Visualização das características dos textos, a execução de tarefas de Classificação e Clusterização de documentos, conforme apresentado em [24], [25], [26], [27] e [5]. A versão do Oracle utilizada neste trabalho foi a 10G.

### 3.1.1. INDEXAÇÃO NO *ORACLE TEXT*

É preciso definir inicialmente que tarefas serão realizadas com os textos. Isso é importante porque irá determinar alguns parâmetros e configurações que deverão ser aplicados aos dados, como os índices.

Existem quatro tipos de índice que podem ser aplicados aos dados textuais para as tarefas de mineração de textos. São eles: *CONTEXT*, *CTXCAT*, *CTXRULE* e *CTXPATH*. Cada um é indicado para determinado tipo de tarefa e são acionados por operadores específicos. Um maior detalhamento destes índices é apresentado em [24].

### 3.1.1.1. O PROCESSO DE INDEXAÇÃO

Será apresentado aqui um sumário do processo de indexação no *OT*, como ilustra a figura 3.2. O processo tem início com a criação dos índices nos dados, seguido da definição dos parâmetros dos procedimentos, que são configurados de acordo com as tarefas a serem realizadas.

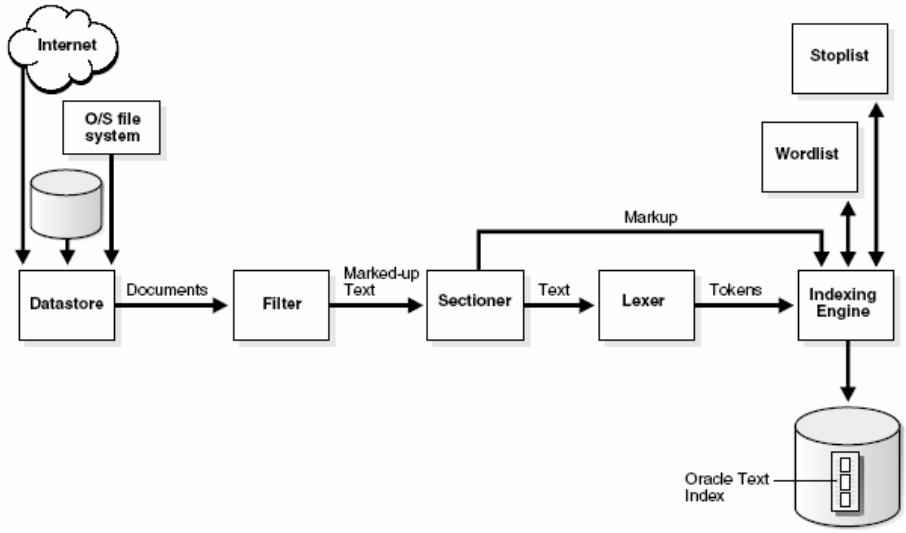


Figura 3.2 : O processo de indexação no *ORACLE TEXT*

A seguir serão descritas as seções envolvidas no processo.

- *Datastore* : O processo tem início quando esse componente lê os documentos, de acordo com o local onde eles se encontram armazenados.

- *Filter* : As ações que esse componente realizará dependem da configuração realizada no parâmetro *FILTER_PREFERENCE*. Caso o parâmetro seja configurado para *NULL_FILTER* ele não executará ação nenhuma. Isso é indicado para documentos no formato de texto plano, *HTML* ou *XML*, que não necessitam de filtragem. Documentos no formato binário devem ser filtrados

automaticamente, caso em que o parâmetro deverá estar configurado para *AUTO_FILTER*. Já documentos escritos em formatos não interpretáveis pelo banco deverão ser convertidos para um formato específico que o banco reconheça, caso em que o parâmetro deverá estar configurado para *CHARSET_FILTER*.

- *Sectioner* : Esse componente separa os dados em texto e informações da seção. O tipo de seção extraído é determinado pelo tipo de grupo de seção configurado. As informações de seção são passadas diretamente para o componente *Indexing Engine*, que fará o seu processamento mais tarde, e o texto segue para o componente *Lexer*.

- *Lexer* : Esse componente decompõe o texto em símbolos, de acordo com a sua linguagem. Esses símbolos normalmente são palavras. Para a extração dos símbolos, ele usa os parâmetros especificados na configuração do atributo *lexer*. Quando a indexação de temas é suportada pela linguagem, esse componente analisa o texto para criar os símbolos dos temas.

- *Indexing Engine* : Esse componente cria um índice invertido que mapeia os símbolos para os documentos que os contém. Se uma *stopword* e uma *stoptheme* forem especificadas, nessa fase o sistema as utiliza para excluir as palavras e temas indicados.

### 3.1.1.2. LOCALIZAÇÃO DO TEXTO

O pré-requisito básico para a execução das tarefas no *OT* é ter uma tabela carregada com os documentos que serão utilizados. Nessa tabela são armazenadas as informações sobre os documentos e é realizada a indexação.

Tarefas de classificação de documentos requerem que seja aplicado à tabela de documentos um índice do tipo *CONTEXT*. Esse tipo de índice é usado em aplicações de recuperação de informações textuais em grandes coleções de documentos. Ele pode ser usado em documentos de diferentes formatos, como no *Microsoft Word, HTML*, PDF ou em textos planos. Como ilustrado na figura 3.3, quando esse tipo de índice é criado, a

tabela de documentos pode ser carregada com algum desses três tipos de elementos, a saber:

- Informação textual (os documentos propriamente ditos);
- Os endereços dos documentos no sistema de arquivos;
- *URLs* que indiquem os endereços dos documentos na Internet.

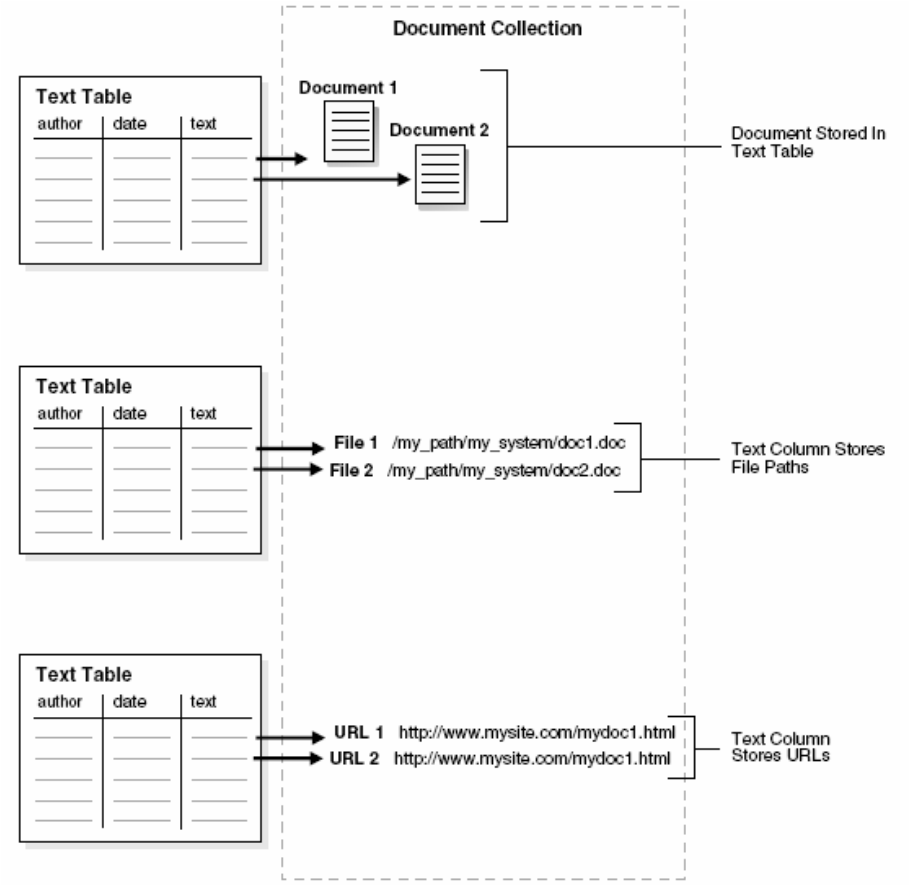


Figura 3.3 : Possíveis opções para o armazenamento dos documentos

### 3.1.1.3. IDIOMA DOS DOCUMENTOS

O *OT* pode indexar documentos em diversos idiomas. Por padrão, o *OT* assume que o idioma do texto seja o mesmo especificado na configuração do banco de dados. Caso não seja, é preciso configurar manualmente o correto.

As preferências do parâmetro *basic_lexer*, que trata das particularidades de cada idioma, podem ser configuradas para indexar, por exemplo, espaços em branco em

documentos em inglês, francês, alemão, holandês e espanhol. Para alguns desses idiomas, outros procedimentos podem ser habilitados, como o tratamento de palavras compostas, o tratamento de palavras coloquiais, fora da norma culta da língua, entre outros. Também podem ser indexados documentos nos idiomas japonês, chinês e coreano. O Oracle 10G não dá suporte para o idioma Português.

### 3.1.1.4. CONSULTAS E INDEXAÇÃO *CASE-SENSITIVE*

Por padrão, inicialmente todos os caracteres textuais são convertidos para caixa alta (maiúsculas) antes da indexação. Isso faz com que as consultas, por padrão, sejam *case-insensitives*, e, portanto, mais flexíveis.

É possível alterar essa configuração padrão e tornar as consultas *case-sensitives*, através da configuração do atributo *mixed_case* do parâmetro *basic_lexer*. Quando se cria um índice *case-sensitive*, é necessário que seja especificado na consulta a grafia exata dos termos dos documentos para que haja sucesso nas buscas. Por exemplo, se um documento contém a palavra *Money* é preciso que na consulta seja especificada a mesma grafia da palavra, ou seja, *Money*. Se for especificado *money* ou *MONEY*, por exemplo, a consulta não retornará o documento.

### 3.1.1.5. CARACTERÍSTICAS ESPECÍFICAS DOS IDIOMA

Algumas características específicas dos idiomas podem ser habilitadas visando dotar o sistema de novas funcionalidades. Lista-se a seguir algumas delas:

- INDEXAÇÃO DE TEMAS:

Nos idiomas inglês e francês é possível, por padrão, a indexação de documentos para a pesquisa de temas. Um tema deve ser um conceito suficientemente desenvolvido ou abordado em um documento. Temas podem ser consultados através do operador *ABOUT*.

É possível a indexação de temas em outros idiomas, desde que sejam executados e compilados no sistema bases de conhecimentos nesses idiomas.

A indexação de temas pode ser habilitada e desabilitada através do atributo *index_themes* do parâmetro *basic_lexer*.

- *STOPWORDS* E *STOPTHEMES*

*Stopwords* são palavras com pouca ou nenhuma capacidade de predição, conforme detalhado no item 2.1.2, que não são indexadas e, portanto, não consideradas, nas tarefas de mineração de textos.

Por padrão, o *OT* disponibiliza uma lista de *stopwords* chamada *stoplist*, para indexação, nos idiomas suportados. É possível alterar essa lista fornecida ou criar outras, e configurar qual lista deve ser utilizada pelo sistema. Também podem ser criadas *stoplists* com termos de vários idiomas para serem utilizadas em tarefas que envolvam documentos em mais de um idioma.

*Stopthemes* são palavras que não devem ser utilizadas na definição nem na nomeação de temas. Listas de *stopthemes* podem ser adicionadas ao sistema através de rotinas especiais e pode-se configurar que lista deve ser considerada pelo sistema. Para essas tarefas deve-se utilizar o pacote *CTX_DDL*. Para maiores detalhes, consultar [25].

- VARIAÇÕES NA GRAFIA DOS TERMOS

Idiomas como alemão, dinamarquês e sueco contêm palavras que possuem mais de uma grafia. Por exemplo, em alemão, *ae* pode ser substituído por *ä*. A grafia *ae* é conhecida como forma alternativa de escrita.

Por padrão, o *OT* indexa termos na sua forma alternativa nesses idiomas. Como conseqüência, essas palavras podem ser consultadas estando escritas em quaisquer formas.

Essa indexação pode ser habilitada e desabilitada através do atributo *alternate_spelling* do parâmetro *basic_lexer*.

- EQUIVALÊNCIA *FUZZY* E *STEMMING*

Equivalência *fuzzy* habilita a consulta de termos escritos de forma equivalente ou similar a grafia correta. Já o *stemming* habilita a consulta de termos com a mesma raiz lingüística. Por exemplo, uma consulta com a palavra *speak* seria expandida para todos os documentos que contiverem *speak*, *speaks*, *spoke* e *spoken*.

Essas duas funcionalidades são automaticamente habilitadas se o *OT* suporta essas características para o idioma especificado como padrão no sistema.

A equivalência *fuzzy* é habilitada com o parâmetro padrão configurado para que a contagem de similaridade seja a maior possível e para o maior número de termos

expandidos ou derivados. Esses parâmetros também podem ser configurados manualmente para os valores que se desejar.

Também podem ser criados índices de *stemming* para um aumento do desempenho do sistema, através da configuração do atributo *index_stems* do parâmetro *basic_lexer*. Uma maior detalhamento dessas características pode ser encontrado em [25].

- CONSULTAS NAS SEÇÕES DOS DOCUMENTOS

Em documentos que possuem uma estrutura interna como os do formato *HTML* e *XML*, é possível definir e indexar seções dos documentos. Com a indexação das seções, fica possível estreitar o escopo das consultas especificando a área de busca em determinadas seções dos documentos. Podem ser definidas seções prioritárias para a indexação e também podem ser especificados grupos de seções preferenciais para as buscas nas consultas. É possível configurar o sistema para que ele automaticamente crie seções em documentos *XML* durante a indexação.

## 3.1.2. CONSULTAS NO *ORACLE TEXT*

As tarefas de consulta no *ORACLE TEXT* podem ser divididas em três diferentes abordagens:

- Consultas em coleções de documentos;
- Consultas em catálogos de informação;
- Pesquisa *XML*.

### 3.1.2.1. CONSULTAS EM COLEÇÕES DE DOCUMENTOS

Tarefas de consulta em coleções de documentos habilitam os usuários a realizar pesquisas em coleções de documentos como bibliotecas digitais, *web sites* ou documentos em *datawarehouses*. Essas coleções são tipicamente estáticas, sem alterações ou atualizações significativas a partir da indexação inicial realizada. Os

documentos podem ser de diferentes tamanhos e formatos, como *HTML*, PDF ou *Microsoft Word* e devem ser carregados em tabelas no bando de dados.

Na especificação das consultas podem ser utilizadas palavras ou frases. Podem ser realizadas consultas que combinem logicamente palavras e frases, através de operadores específicos. Outros operadores contendo técnicas de *stemming*, pesquisa por proximidade e utilizando palavras-chave podem ser empregados a fim de aprimorar os resultados das buscas.

Um importante fator neste tipo de aplicação é a recuperação apenas de documentos relevantes para a consulta. A lista de resultados deve ordenar os documentos em ordem crescente de relevância, isto é, dos mais relevantes para os menos relevantes.

### 3.1.2.2. CONSULTAS EM CATÁLOGOS DE INFORMAÇÃO

Catálogos de informação consistem em repositórios de informações que são atualizados regularmente, como em um *site* de uma livraria ou de uma loja virtual, por exemplo.

Este tipo de tarefa normalmente requer a combinação de um componente textual, como o nome de um livro, com um componente estruturado, como o preço do livro. Os resultados das buscas, de uma forma geral, retornam um componente estruturado como o preço ou a data da operação. Um bom tempo de resposta é sempre um requisito importante na avaliação deste tipo de tarefa.

### 3.1.2.3. PESQUISA *XML*

A chamada pesquisa *XML* pode ser definida no *OT* como a consulta em documentos no formato *XML*. Ela apresenta vantagens em relação à busca convencional: em documentos convencionais, as pesquisas são realizadas no documento como um todo; já a pesquisa *XML* pode realizar as buscas em seções específicas dos documentos, como o título ou o corpo, restringindo, assim, a área de busca e otimizando o desempenho do sistema.

O *OT* possibilita a pesquisa *XML* a partir de diferentes técnicas:

- Utilizando *ORACLE TEXT*;
- Utilizando *Oracle XML DB Framework*;
- Combinando as características do *ORACLE TEXT* com o *Oracle XML DB.*

Para maiores informações sobre as abordagens de pesquisa *XML* disponíveis no *OT*, consultar [24].

Neste trabalho, foram realizados experimentos de consulta baseados na abordagem de Pesquisa *XML*. Porém, como a base de dados utilizada nos Estudos de Casos não possibilitava nenhum ganho na execução das tarefas a partir dessa abordagem, além do fato dessa estratégia agregar mais complexidade às tarefas, optou-se pela abordagem baseada em coleções de documentos.

### 3.1.3. CLASSIFICAÇÃO NO *ORACLE TEXT*

O *OT* apresenta diferentes abordagens de Classificação de documentos. Uma tarefa de Classificação executa as suas ação baseada na análise do conteúdo dos documentos, conforme ilustra a figura 3.4. Essa ação pode ser, por exemplo, a associação dos documentos às classes pré-definidas ou o envio de mensagens por e-mail avisando da chegada de novos documentos de determinada categoria. O resultado final da tarefa é um conjunto de documentos categorizados.

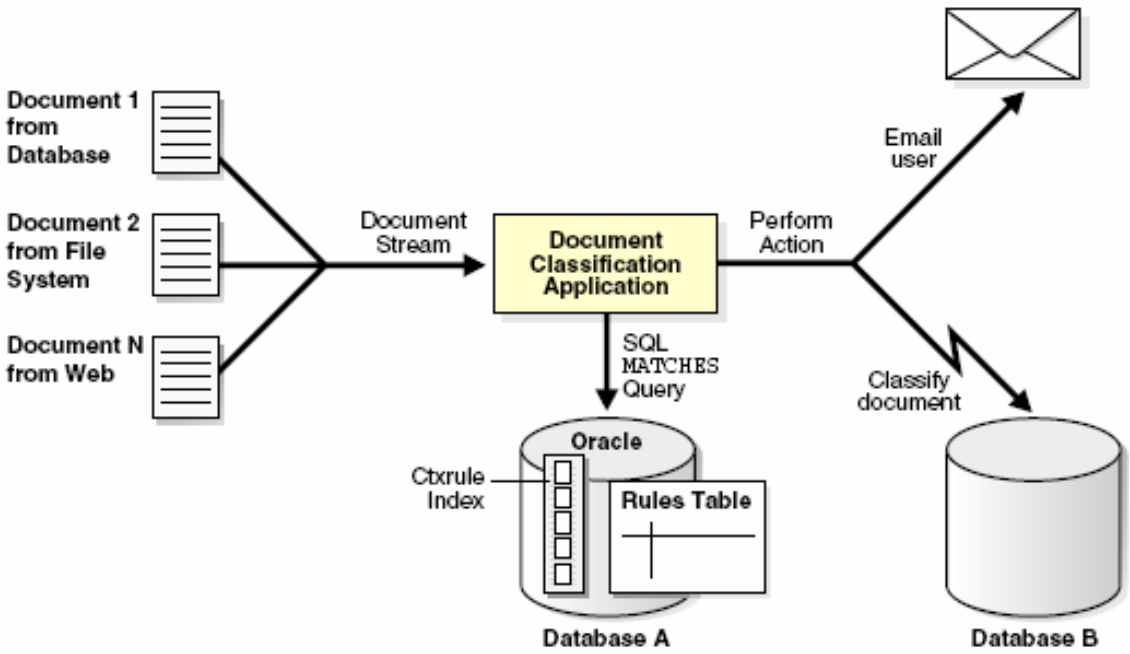


Figura 3.4 : Visão geral das aplicações de Classificação de documentos

Abaixo são apresentadas as abordagens de Classificação de documentos no *OT*:

- Classificação Baseada em Regras;
- Classificação Supervisionada:
  - baseada em Árvores de Decisão (AD);
  - baseada no algoritmo *SVM* (*Support Vector Machine*);

Na Classificação Baseada em Regras, o usuário deve definir as regras de classificação dos documentos. Já na Classificação Supervisionada o *OT* define as regras automaticamente, baseado em um conjunto de documentos pré-classificados fornecido pelo usuário, que integram a base de treinamento.

### 3.1.3.1. CLASSIFICAÇÃO BASEADA EM REGRAS

Classificação baseada em regras é a abordagem básica de Classificação disponibilizada pelo *OT*. Nela o usuário deve alimentar todo o sistema para que a tarefa seja executada. Os principais passos para isso são os seguintes:

1- Criação de uma tabela para conter os documentos a serem classificados, carga de dados e indexação da tabela.
2- Criação de uma tabela com a descrição das categorias e a definição das regras de cada uma. Essas regras basicamente são termos ou expressões a serem buscados nos documentos. Por exemplo, poderia ser criada uma categoria chamada ‘Medicina’ e definido a regra de que os documentos que contivessem os termos ‘hospital’ ou ‘infecção’ ou ‘medicina’ seriam da categoria ‘Medicina’.
3- Indexação da tabela de categorias com um índice do tipo *CTXRULE*.
4- Criação da tabela de resultados, que será carregada através da execução da *procedure* de classificação e conterá os documentos associados às categorias.
5- Configuração da *procedure* e execução da mesma.

Após a execução dos passos descritos acima, a tabela de resultados é automaticamente carregada e o resultado da classificação poderá, então, ser visualizado.

VANTAGENS:

- A possibilidade do usuário definir e manipular as regras de classificação. Isso permite, entre outras coisas, que o conhecimento de um especialista possa ser incorporado ao sistema.
- Para pequenas e médias coleções de documentos os resultados costumam ter uma acurácia alta, o que depende, obviamente, da qualidade das regras geradas.

DESVANTAGENS:

- A definição das regras para amostras grandes de documentos e com muitas categorias torna-se uma tarefa complexa e dependente de um especialista.
- Se a amostra inicial de documentos for acrescida de novos documentos e de novas categorias, será preciso revisar e completar manualmente as regras definidas inicialmente.

### 3.1.3.2. CLASSIFICAÇÃO SUPERVISIONADA

Na Classificação Supervisionada, é utilizada a *procedure CTX_CLS.TRAIN* para a geração automática das regras de Classificação a partir de uma base de treinamento

fornecida. Essa é a maior vantagem dessa abordagem em relação à Classificação baseada em regras. Porém, antes da execução da *procedure*, também deverá ser criada manualmente pelo usuário uma tabela de categorias, contendo somente a descrição das categorias (não mais as regras, que serão automaticamente geradas).

Após a execução da *procedure* e da geração automática das regras, a tabela de categorias deverá ser indexada com um índice do tipo *CTXRULE*. A partir de então, poderá ser realizada a Classificação de novos documentos.

VANTAGENS:

- As regras de classificação são escritas automaticamente a partir da base de treinamento fornecida.
- Para grandes coleções de documentos com várias categorias, essa geração automática de regras é muito útil.

DESVANTAGENS:

- É necessário que se tenha uma amostragem razoável de documentos nas categorias que se deseja tratar para alimentar a base de treinamento.
- Regras geradas automaticamente podem não ser tão específicas ou precisas quanto aquelas geradas com conhecimento de especialistas.

O *OT* disponibiliza duas abordagens de Classificação Supervisionada, baseadas em algoritmos distintos. São elas:

- Classificação Supervisionada baseada em Árvores de Decisão;
- Classificação Supervisionada baseada em no algoritmo *SVM* (*Support Vector Machine*).

### 3.1.3.2.1. CLASSIFICAÇÃO BASEADA EM ÁRVORES DE DECISÃO

Nessa abordagem é necessário configurar o parâmetro da *procedure CTX_CLS.TRAIN* para *RULE_CLASSIFIER*. Ela utiliza um algoritmo baseado em árvores de decisão para a geração das regras. Em linhas gerais, uma árvore de decisão é um método que decide entre dois nós (ou mais, mas tipicamente dois). Nesse caso, os

nós são: '*esse documento pertence a alguma categoria dada*' ou '*esse documento não pertence a nenhuma categoria dada*', integrante da base de treinamento.

Esse algoritmo possibilita o teste de um conjunto de atributos, entre eles:

- Termos dos documentos;
- Raízes dos termos dos documentos;
- Temas do documento (se esse atributo for suportado pelo idioma em uso).

O algoritmo de treinamento constrói uma ou mais árvores de decisão para cada categoria descrita na base de treinamento.

As árvores de decisão incluem o conceito de confiança. A cada regra gerada é alocado um valor percentual que representa a precisão da regra, determinado a partir da base de treinamento. Em casos triviais, essa precisão é quase sempre de 100%, mas isso somente representa as limitações da base de treinamento. Nesse caso, as regras geradas podem parecer simples demais, mas elas geralmente são suficientes para distinguir as categorias da base de treinamento.

Os principais passos para a configuração do sistema pelo usuário são os seguintes:

1- Criação da tabela de documentos da base de treinamento do algoritmo para a geração do classificador, carregamento e indexação da mesma.
2- Criação e carregamento de uma tabela com a descrição das categorias.
3- Criação da tabela de treinamento, onde os documentos são associados às categorias, carregamento e indexação da mesma.
4- Criação da tabela de regras, que será carregada com a execução da *procedure* de geração de regras e conterá as regras geradas automaticamente pela mesma.
5- Configuração da *procedure* de geração das regras a partir da base de treinamento e execução da mesma.
6- Indexação da tabela de regras com um índice do tipo *CTXRULE*.
7- Configuração da *procedure* de classificação e execução da mesma. Cabe ressaltar aqui que o documento a ser Classificado é um dos parâmetros a ser configurado na *procedure* e que esse procedimento é realizado individualmente para cada documento.

A maior vantagem dessa abordagem é que as regras geradas são observáveis e editáveis pelo usuário. Já a maior desvantagem é configurar e executar uma *procedure* para cada documento a ser classificado.

Essa técnica se aplica quando se deseja, por exemplo, que o sistema gere um conjunto inicial de regras, que posteriormente serão analisadas, refinadas e complementadas por um especialista.

### 3.1.3.2.2. CLASSIFICAÇÃO BASEADA NO ALGORITMO *SVM*

Essa é a segunda abordagem de Classificação Supervisionada disponibilizada pelo *OT*, conhecida como Classificação baseada em *SVM* (*Support Vector Machine*). *SVM* é um algoritmo de aprendizado de máquina derivado da teoria de aprendizagem estatística. Um maior detalhamento desse algoritmo pode ser encontrado em [14] e [28].

Uma propriedade importante dessa abordagem é a sua capacidade de extrair regras de Classificação a partir de uma base muito pequena de treinamento.

O uso desse método é muito similar ao uso do método baseado em Árvores de Decisão, com algumas diferenças em relação à configuração de parâmetros na execução das *procedures*, além de outras na definição da tabela de categorias e na de resultados.

Os principais passos para a configuração do sistema pelo usuário são os seguintes:

1- Criação da tabela de documentos da base de treinamento do algoritmo para a geração do classificador, carregamento e indexação da mesma.
2- Criação da tabela de treinamento, que conterá a descrição das categorias e a associação dos documentos a elas, carregamento e indexação da mesma.
3- Criação da tabela de regras, que será carregada com a execução da *procedure* e conterá as regras geradas automaticamente pela mesma.
4- Configuração da *procedure* de geração das regras a partir da base de treinamento e execução da mesma.
5- Indexação da tabela de regras com um índice do tipo *CTXRULE*.
6- Configuração da *procedure* de Classificação e execução da mesma. Cabe ressaltar aqui que esse procedimento é realizado individualmente para cada documento a ser classificado.

Esse método gera regras de classificação binárias, que são opacas para o usuário, ou seja, são visíveis mas não são interpretáveis, e, portanto, não são editáveis ou passíveis de complementação, como são as regras geradas pelo método baseado em árvores de decisão. Por outro lado, esse método frequentemente apresenta resultados de maior precisão que os encontrados com a outra abordagem de Classificação Supervisionada.

Para a execução desse método, a memória alocada deverá ser suficientemente grande para a carga completa do modelo *SVM*. Caso contrário, não será possível executar as rotinas, pois a aplicação incorrerá em sucessivos erros de falta de memória.

### 3.1.4. CLUSTERIZAÇÃO NO *ORACLE TEXT*

Na Clusterização o usuário não precisa especificar nenhuma regra de definição dos grupos nem fornecer uma base de treinamento ao sistema, basta configurar os parâmetros que o sistema gera automaticamente as regras, cria os grupos e associa os documentos a eles.

Uma vez configurado os parâmetros, a realização dessa tarefa ocorre de forma automática pela *procedure CTX_CLS.CLUSTERING*. O *OT* analisa estatisticamente o conjunto de documentos fornecido e, de acordo com os seus conteúdos, os relaciona aos grupos encontrados.

Essa *procedure* cria uma hierarquia de grupos de documentos, conhecidos como *Clusters*, e, para cada documento, retorna uma pontuação de relevância para cada um dos *Clusters* encontrados. A partir daí, é necessário fazer uma análise manual dessas pontuações para poder associar cada documento ao *Cluster* para o qual se obteve a maior contagem.

Semelhante ao observado na Classificação Supervisionada baseada em árvores de decisão, os atributos utilizados para a determinação dos *Clusters* podem consistir em termos dos documentos, raízes dos termos dos documentos ou temas do documento (caso esse atributo seja suportado pelo idioma em uso).

Os principais passos para a configuração do sistema pelo usuário são os seguintes:

1- Criação de uma tabela para conter os documentos a serem agrupados, carregamento e indexação da mesma.
2- Criação de 2 tabelas distintas de resultados, que serão carregadas com a execução da *procedure*, a saber:
   - Uma tabela de descrição dos *Clusters*, que conterá informações sobre a sua geração, como a identificação, os principais termos encontrados, uma sugestão de rótulo e a contagem da qualidade do *Cluster*.
   - Uma tabela mostrando as pontuações das associações de cada documento a todos os *Clusters* gerados. Avaliando-se essas informações, é possível associar cada documento ao *Cluster* que obteve a maior pontuação.
3- Configuração da *procedure* de Clusterização e execução da mesma.

A *procedure CTX_CLS.CLUSTERING* utiliza o algoritmo *K-MEANS* para executar o agrupamento dos documentos. Ela pode ser configurada através do parâmetro *KMEAN_CLUSTERING*, onde é possível se definir o valor do *K*, que determinará o número de *Clusters* a ser encontrado.

VANTAGENS:

- Não é necessário fornecer ao sistema regras de categorização ou mesmo bases de treinamento com conjuntos pré-categorizados de documentos.
- Possibilita a descoberta de padrões e semelhanças de conteúdo nos documentos não suspeitadas ou esperadas pelos usuários.

DESVANTAGENS:

- Esse método pode resultar em *Clusters* inconsistentes, já que as operações não são configuradas pelos usuários, mas baseadas em algoritmos internos.
- As regras que definem os *Clusters* geradas automaticamente não são visíveis para os usuários.
- Essa é uma técnica que requer muito processamento computacional e, dependendo do tamanho do conjunto de documentos ou dos recursos computacionais disponíveis, pode requerer bastante tempo ou mesmo se mostrar inviável.

Esse método pode ser utilizado quando não se tem uma idéia clara das regras ou dos grupos existentes no conjunto de documentos. Uma possível abordagem é utilizar esse método para gerar um conjunto inicial de categorias, e, na sequência, utilizar algum método de Classificação Supervisionada para refinar essa categorização inicial.

### 3.1.5. VISUALIZAÇÃO NO *ORACLE TEXT*

Nas consultas no *OT*, os documentos ou trechos de documentos retornados podem ser apresentados com os termos da consulta em destaque, nos casos das consultas de conteúdo, ou com os temas em destaque, no caso das consultas de temas.

Podem ser gerados três tipos de resultado associados ao destaque de termos ou temas. São eles:

- Uma versão do documento com marcações nos termos da consulta;
- Uma versão do documento com termos ou trechos dos documentos em destaque;
- Uma apresentação do elemento consultado dentro do seu contexto.

As *procedures* que tratam das consultas com destaque dos termos ou temas integram o pacote *CTX_DOC* do *OT*.

O *OT* disponibiliza outras funcionalidades relacionadas com a visualização dos documentos. São as seguintes:

- Obtenção do Assunto Principal do documento;
- Obtenção da Lista de Temas do documento;

#### 3.1.5.1. VERSÃO COM MARCAÇÕES NOS TERMOS DA CONSULTA

Essa tarefa pode ser realizada pelas *procedures CTX_DOC.MARKUP* ou *CTX_DOC.POLICY_MARKUP* (que são equivalentes, com a diferença de que a segunda não requer índice), que são responsáveis por recuperar a referência do

documento indicado e apresentar o resultado com uma versão do documento com marcações nos termos da consulta.

Esse resultado pode ser apresentada de duas formas:

- Através de um texto plano com os termos marcados com os símbolos ‘<’ e ‘>’ (por exemplo, <<<*Financial*>>> );
- Com os termos com marcações no padrão *HTML* (por exemplo, <*b*>*Financial*</*b*>).

### 3.1.5.2. VERSÃO COM TERMOS EM DESTAQUE

Para a realização dessa tarefa podem ser usadas as *procedures CTX_DOC.HIGHLIGHT* ou *CTX_DOC.POLICY_ HIGHLIGHT*.

De forma semelhante à tarefa apresentada no item 3.1.5.1, o resultado com os termos ou trechos dos documentos em destaque pode ser apresentado em um texto plano ou no formato *HTML*.

O destaque dos termos pode ser configurado pelo usuário de acordo com os seguintes padrões:

- Uma fonte diferente da original (por exemplo, <<<*Financial* >>> , no caso de texto plano);
- Uma cor diferente da original (por exemplo, <*b*>*Financial*</*b*> , no caso do formato *HTML*);
- Um tamanho diferente do original (por exemplo, <*b*>*Financial*</*b*> , no caso do formato *HTML*).

### 3.1.5.3. APRESENTAÇÃO COM ELEMENTO CONTEXTUALIZADO

Essa tarefa pode ser realizada pelas *procedures CTX_DOC.SNIPPET* ou *CTX_DOC.POLICY_ SNIPPET*. Como resultado da consulta é apresentado o termo ou o trecho do documento recuperado dentro do contexto em que aparece no documento.

Esse resultado também é conhecido com *Key Word in Context* ou *KWIC*, porque, ao invés de retornar o documento inteiro (com ou sem destaque nos termos da consulta), ele retorna um fragmento de texto, permitindo ao usuário observar o termo no seu contexto.

O usuário pode configurar o destaque a ser dado nos termos da consulta dentro do fragmento de texto retornado, de acordo com os destaques previstos na tarefa apresentada no item 3.1.5.2.

### 3.1.5.4. OBTENÇÃO DO ASSUNTO PRINCIPAL DO DOCUMENTO

O assunto principal de um documento é o termo ou a expressão que melhor descreve sobre o que trata o documento.

Para a sua obtenção, é utilizada a *procedure CTX_DOC.GIST*.

O resultado da obtenção do assunto principal pode ficar armazenado em memória e ser exibido para o usuário logo após a consulta ou ser armazenado em uma tabela previamente criada para essa finalidade. O usuário deve configurar o formato desejado.

Pode ser especificado pelo usuário o tamanho do termo ou da expressão a ser retornada pela consulta, na configuração dos parâmetros da *procedure*.

### 3.1.5.5. OBTENÇÃO DA LISTA DE TEMAS DO DOCUMENTO

Uma lista de temas é uma lista dos conceitos principais encontrados em um documento.

Para a sua obtenção, deve ser utilizada a *procedure CTX_DOC.THEMES*.

Os resultados da geração da lista também podem ficar armazenados somente em memória e ser exibidos para o usuário logo após a consulta ou podem ser armazenados em uma tabela previamente criada para essa finalidade. O usuário é responsável por configurar o formato desejado.

Pode ser obtida uma lista simples, onde cada tema encerra um conceito isolado, ou uma lista hierárquica, onde os temas podem estar relacionados entre si.

### 3.1.6. TRABALHANDO COM O *THESAURUS* NO *ORACLE TEXT*

Usuários de aplicações de consulta, ao procurar por informações de um determinado tópico, podem não ter um conhecimento preciso dos termos que foram usados nos documentos referentes ao dado tópico. Isso pode atrapalhar ou mesmo inviabilizar a consulta. A utilização de um dicionário de termos ou *Thesaurus* pode vir a sanar essa dificuldade.

O *OT* possibilita a criação de dicionários do tipo *Thesaurus*, *case-sensitives* ou não, que definam sinônimos e relações hierárquicas entre termos e frases.

O uso de um *Thesaurus* permite que as tarefas recuperem documentos que contenham informações relevantes, através da expansão das consultas, ao levar em consideração, também, termos similares ou relacionados aos termos originais, de acordo com as definições expressas no dicionário.

Vários dicionários desse tipo podem ser criados e o usuário pode configurar qual dicionário vai ser usado para que tarefa.

O *OT* permite a criação e manipulação de *Thesaurus* através de dois pacotes de funcionalidades: o *CTXLOAD* e o *CTX_THES.CREATE_THESAURUS*. Para maiores detalhes, consultar [25].

Funcionalidades específicas podem ser acionadas para carregar um *THESAURUS* a partir de uma lista de termos em um texto plano, por exemplo. Também é possível exportar dicionários carregados no sistema.

Se um *Thesaurus* não for especificado durante uma consulta, por padrão o sistema chama um denominado *Default*. Entretanto, o sistema não fornece um *Thesaurus Default*. Com isso, se houver interesse em trabalhar com um dicionário padrão nas operações que envolvam o *Thesaurus*, é preciso carregar um no sistema com o nome *Default*.

Ainda que o *OT* não forneça um *Thesaurus Default*, ele pode suprir essa lacuna, fornecendo um arquivo no formato *ctxload*, que pode ser usado para criar um dicionário geral na língua inglesa. E esse dicionário geral pode, então, ser usado para a criação de um *Thesaurus Default*.

## 3.2 TOAD

O Toad é um software comercial de gerenciamento de plataformas de banco de dados, desenvolvido pela empresa Quest Software. Existem versões distintas para as plataformas Oracle, SQL Server, DB2 e MySQL.

Foi utilizada a versão 8.0.0.47 para o Oracle, que dá suporte para as versões 8, 9 e 10 desse SGBD, conforme apresentado em [29] e ilustrado pela figura 3.5.

Ele funciona como uma espécie de interface do SGDB, sendo a sua utilização bastante amigável, tornando as tarefas de criação e manutenção das tabelas, de edição de registros, criação e alteração de índices, configuração dos parâmetros, criação e execução de *procedures* e visualização dos resultados, pra citar algumas, muito mais práticas e transparentes para o usuário.

Diferentemente do *SQL Plus* do Oracle, onde a entrada de instruções se dá através de linhas de comando no terminal, no Toad é possível passar vários tipos de instrução através do uso de menus e de botões, o que torna as tarefas bem mais simples.

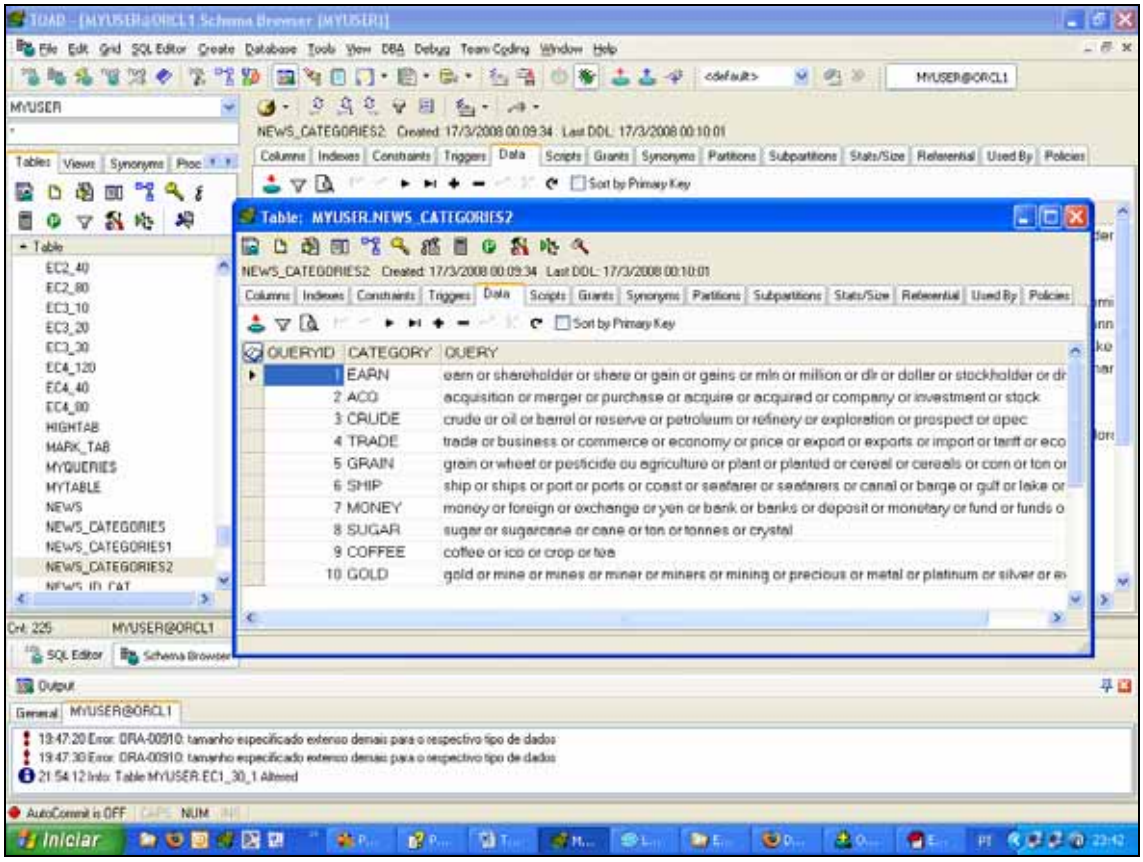


Figura 3.5 : A interface do Toad aplicado ao Oracle

## 3.3 *POLY ANALYST*

*Poly Analyst* (*PA*) é um software desenvolvido pela empresa *Megaputer Intelligence Inc.*, tendo sua primeira versão lançada em 1997. Ele é um programa de mineração de dados (estruturados e não estruturados) com variados propósitos, visando facilitar as pesquisas e a tomada de decisões a partir da análise dos dados, conforme apresentado em [30] e ilustrado pela figura 3.6.

Ele contém várias ferramentas para exploração dos dados, que possibilitam ligá-los, alterá-los, separá-los, analisá-los e sumarizá-los. Elas estão disponíveis para que o usuário possa compor seus projetos de análise de forma personalizada, de acordo com as tarefas que pretende realizar.

De forma resumida, listamos abaixo algumas tarefas a serem realizadas com o apoio do programa:

- Importar bases de dados de diferentes origens;
- Mesclar diferentes bases de dados;
- Adicionar, alterar ou remover colunas ou registros;
- Filtrar, eliminar ou tratar valores ausentes ou com ruído;
- Excluir dados duplicados;
- Adicionar informações sobre a base de dados;
- Visualizar estatísticas básicas a respeito dos dados;
- Realizar tarefas como a descoberta de Associações, Classificação, Clusterização e Predição com os dados;
- Analisar dados textuais e encontrar palavras-chave;
- Visualizar os dados ou o resultado das análises de forma personalizada;
- Exportar bases de dados;
- Gerar relatórios com os resultados das análises realizadas;

O *PolyAnalyst* é um pacote de software intuitivo com um ciclo de aprendizagem curto. Ele pode ser compreendido e operado em um nível mais simples ou em nível avançado. No nível mais simples ele requer somente conhecimentos básicos em computação, em conceitos de bancos de dados e em estatística. Já no nível avançado são

requeridos conhecimentos em lógica de programação, conceitos básicos de Inteligência Artificial e noções avançadas de estatística.

Neste trabalho foi utilizada a versão 6 do programa.

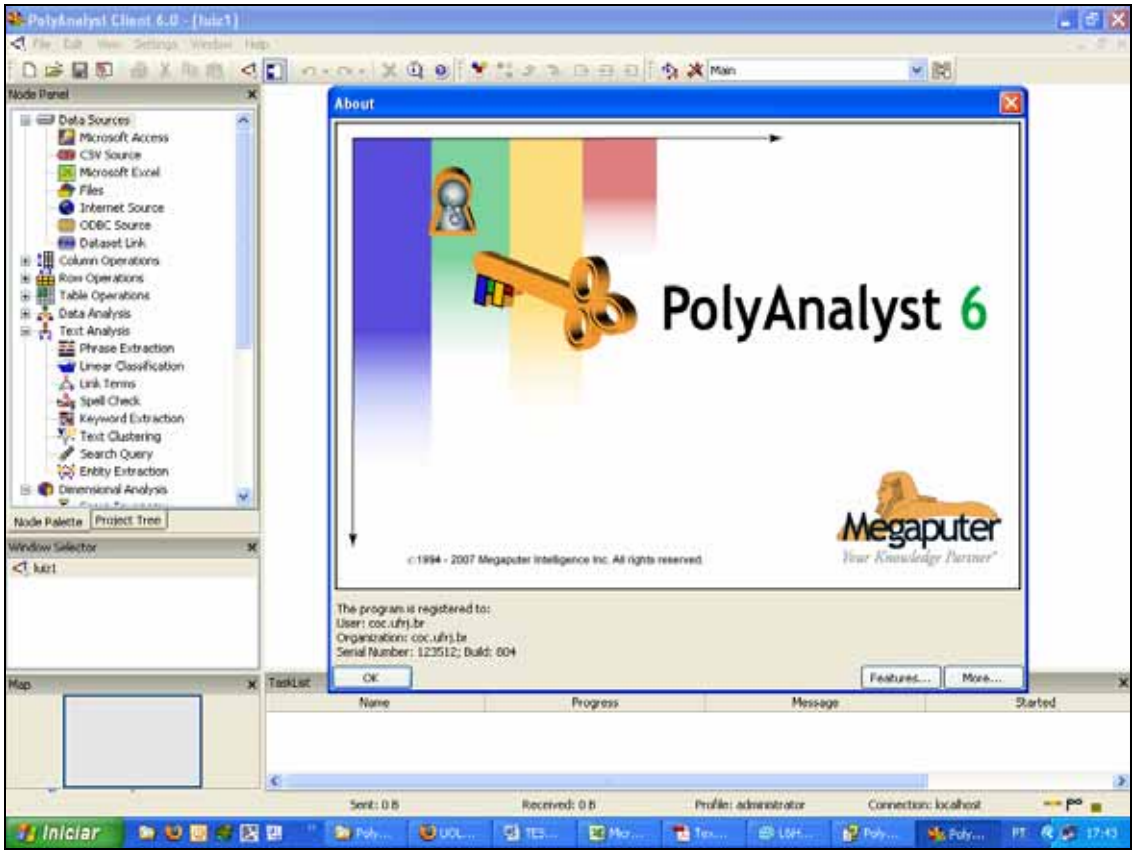


Figura 3.6 : Tela inicial do programa *Poly Analyst*

### 3.3.1. A ARQUITETURA DO *POLY ANALYST*

O software utiliza a arquitetura cliente/servidor, que permite uma administração eficiente dos recursos disponíveis. Tipicamente são montadas várias estações clientes, que são atendidas por um único servidor.

Essas estações podem ser configuradas individualmente para dar suporte a tarefas específicas. Por exemplo, algumas podem estar configuradas para as tarefas de análises de dados propriamente ditas, enquanto outras serem voltadas para usuários de negócio, que, basicamente, terão interesse nos relatórios das análises realizadas.

O software é composto por diferentes programas, que são listados abaixo, com as suas descrições:

- ***Analytical Client*** – utilizado para a criação e execução dos projetos de análises de dados;
- ***Report Editor*** – utilizado para a geração de relatórios a partir dos resultados das análises.
- ***Dashboard*** – utilizado para usuários de negócio visualizarem e manipularem os relatórios gerados;
- ***Administrative Client*** – utilizado por usuários administrativos para o gerenciamento do software;
- ***Command Line Client*** – utilizado para desenvolver projetos em grupo.
- ***Poly Analyst Server*** – utilizado como servidor de aplicação do sistema.

Neste trabalho foi utilizado o programa ***Analytical Client*** para a execução das tarefas apresentadas, além do servidor (***Poly Analyst Server***).

O ***Analytical Client***, por sua vez, possui diferentes nós para o tratamento dos dados. Esses nós são divididos de acordo com as características das tarefas a serem realizadas. Abaixo temos a apresentação de cada um deles:

- ***Data Sources*** – dá suporte as diferentes fontes de dados passíveis de uso pelo sistema;
- ***Column Operations*** – dá suporte as tarefas de tratamento de colunas;
- ***Row Operations*** – dá suporte as tarefas de tratamento de registros;
- ***Table Operations*** – dá suporte às tarefas que envolvem operações nas tabelas;
- ***Data Analysis*** – dá suporte as tarefas de análise de dados convencionais;
- ***Text Analysis*** – dá suporte as tarefas de análise de dados textuais;
- ***Dimensional Analysis*** – dá suporte as tarefas de segmentação de tabelas;
- ***Charts*** – dá suporte as tarefas de geração de gráficos para a visualização dos resultados;

Neste trabalho foram utilizados os nós ***Data Sources*** e ***Text Analysis*** para a realização das tarefas propostas. O nó ***Charts*** não foi utilizado pois os resultados que ele produz não encontram nenhuma equivalência com os resultados produzidos pelo

*ORACLE TEXT*, e o principal intuito do uso do *Poly Analyst* é validar os resultados obtidos com o Oracle.

A seguir temos as descrições dos nós utilizados.

### 3.3.2. O NÓ *DATA SOURCES*

O sistema permite que se trabalhe com arquivos de diferentes formatos para a entrada de dados, a saber:

- Formato *CSV* (*Comma-Separated Values*);
- Formatos oriundos dos seguintes bancos de dados:
  - Oracle
  - Microsoft SQL Server
  - Microsoft Access
  - Microsoft Excel
  - IBM DB2
  - MySQL
  - Informix
  - Visual FoxPro
  - DBF ou dBASE
- Formato de *webpages* (arquivos carregados diretamente da Internet);
- Formatos de texto diversos, para serem processados em rotinas de análises de texto;
- Formato interno oriundo de um outro nó.

Esse nó tem o propósito de tratar os diferentes tipos de fonte de dados possíveis de alimentar o sistema.

Nele é possível configurar os tipos de dados das colunas das tabelas, indexar os campos, entre outras tarefas. Ele permite, também, que se configure, por exemplo, que numa dada tabela, um determinado conjunto de registros deverá ser considerado válido para as tarefas subsequentes, em detrimento de outros.

### 3.3.3. O NÓ *TEXT ANALYSIS*

Esse nó tratadas tarefas de processamento dos dados textuais, dando suporte a Classificação e a Clusterização de documentos, bem como a extração de frases, palavras-chave ou entidades contidas nos textos, assim como a tarefas de tratamento de erros. Os resultados gerados consistem em relatórios e tabelas que exibem o conteúdo extraído.

O programa possibilita que sejam utilizados um ou mais dicionários no apoio as tarefas de análise de dados textuais. Ele fornece um dicionário padrão para isso, denominado *WorldNet*, um dicionário popular disponível na *web*, que é um modelo bastante abrangente e que normalmente auxilia de forma satisfatória as tarefas de análise de textos de temas gerais, e permite que os usuários cadastrem outros dicionários no sistema para a mesma finalidade.

O *Poly Analyst* incorpora a técnica de *stemming* para o processamento de dados textuais. Ele utiliza um algoritmo próprio desenvolvido para esse fim, baseado em algoritmos conhecidos e disponíveis no mercado. Ele utiliza também um dicionário base específico para o algoritmo de *stemming*, na língua inglesa, contendo diversos termos e o relacionamento entre eles. Esse dicionário base é uma forma modificada do *WorldNet*.

No cálculo de relevância dos termos nos documentos é utilizado um algoritmo para o contagem da pontuação desses termos, que é uma variação do algoritmo *Vector Space Relevance*, conhecido algoritmo desenvolvido com essa finalidade. Uma das alterações introduzidas no novo algoritmo é a agregação do cálculo de proximidade com os demais termos.

Esse nó é composto por oito módulos, responsáveis por tarefas distintas, conforme o apresentado a seguir:

- *Linear Classification* – utilizado para o desenvolvimento de tarefas de Classificação, através de dois algoritmos distintos: o *SVM* e o Bayesiano Simples;
- *Text Clustering* – utilizado para tarefas de Clusterização, a partir de palavras-chave contidas nos textos;

- *Phrase Extraction* – utilizado para extrair automaticamente frases que ocorrem nos textos;
- *Keyword Extraction* - utilizado para identificar palavras-chave e frases;
- *Entity Extraction* - utilizado para extrair automaticamente entidades que ocorrem nos textos, como, por exemplo, nomes e lugares;
- *Search Query* – utilizado para a pesquisa de termos e frases;
- *Link Terms* – utilizado para identificar associações entre termos e frases;
- *Spell Check* – utilizado para tratar erros na grafia dos termos;

A seguir será detalhado o funcionamento dos módulos utilizados neste trabalho, o *Linear Classification* e o *Text Clustering*.

#### 3.3.3.1. *LINEAR CLASSIFICATION*

Esse nó desenvolve um modelo de Classificação dependente de um atributo estruturado, através da utilização de uma coluna independente com os textos. Esse modelo é baseado na frequência e na distribuição dos termos no texto. A partir disso, o programa treina um modelo para a classificação automática de textos.

O *Poly Analyst* apresenta duas abordagens de classificação de textos, baseadas em algoritmos distintos, a saber:

- Baseada no algoritmo *SVM* – um algoritmo que requer um processamento mais intensivo em termos computacionais e que tipicamente apresenta uma maior acurácia nos resultados;
- Baseada no algoritmo Bayesiano Simples – um algoritmo de processamento computacional mais rápido, que é mais escalável, e que, em geral, obtém resultados menos precisos que o *SVM*.

Os detalhes de como os dois algoritmos trabalham não são apresentados na documentação disponibilizada pelo fabricante.

Na configuração dos parâmetros para se executar a Classificação é necessário definir o atributo a ser utilizado como fonte dos dados, o algoritmo de processamento, o

uso ou não de uma *stoplist* e a definição da mesma, o tipo de dado a ser tratado, entre outros.

Durante o processamento, o programa armazena as palavras-chave de forma booleana em uma tabela, indicando se elas aparecem ou não em cada documento, e a frequência com que isso acontece.

Após o processamento, o resultado é apresentado em duas abas: uma contendo as informações da configuração adotada e a outra contendo as informações da classificação propriamente dita, incluindo uma matriz com as taxas de erro por classe.

#### 3.3.3.2. *TEXT CLUSTERING*

Esse nó é utilizado para a Clusterização de documentos. Para a geração dos *Clusters*, ele utiliza uma variação do algoritmo *Suffix Tree Clustering*, que foi apresentado originalmente em [31].

Esse algoritmo apresenta algumas qualidades, entre elas a sua velocidade de processamento, que é próxima de um processamento linear, ou seja, o tempo é proporcional ao número de registros. Ele apresenta resultados de fácil interpretação por parte dos usuários. Também possui mecanismos de identificação das frases.

As frases apresentam a vantagem de ter um poder descritivo mais alto que os termos isolados. Daí, elas se prestam melhor para descrever o conteúdo dos grupos para os usuários, e de uma maneira mais concisa.

O processo envolve dois passos principais: no primeiro, o algoritmo faz uma busca por registros que compartilham frases; no segundo, ele agrupa os documentos a partir da frequência da ocorrência dessas frases.

Na configuração dos parâmetros para a execução da Clusterização primeiro é preciso definir o atributo a ser utilizado como fonte de dados, a partir da tabela de entrada especificada. Então, é necessário configurar alguns parâmetros matemáticos que manipulam o comportamento do algoritmo, como os apresentados a seguir:

- Escolher se o agrupamento a ser feito deverá ser base, exclusivo ou hierárquico;

- O número máximo de agrupamentos básicos, que diz respeito ao número máximo de frases individuais e palavras que serão buscadas, no primeiro passo do algoritmo;
- Os números ou percentuais mínimo e máximo de registros por grupo;
- O uso ou não de um *thesaurus*;
- O uso ou não de um dicionário e a definição de qual.

O resultado do processamento é apresentado em várias abas, contendo a configuração adotada no processamento, um gráfico estatístico com a distribuição dos documentos pelos *Clusters*, uma tabela com os *Clusters* gerados, contendo as suas descrições, a quantidade e a identificação dos documentos presentes, um gráfico mostrando a proximidade dos *Clusters*, entre outros.

Cabe ressaltar que na abordagem do algoritmo implementado para essa tarefa não é possível definir a quantidade de grupos a ser encontrada, na configuração dos parâmetros.

# Capítulo IV

# ESTUDOS DE CASOS

Neste capítulo será descrita inicialmente a base de dados utilizada, suas características e as opções que foram feitas para o desenvolvimento desse trabalho. A seguir serão apresentadas as tarefas realizadas e por último os Estudos de Casos (EC) propriamente ditos, com informações sobre as configurações adotadas, os resultados obtidos e a análise deles.

## 4.1 A BASE DE DADOS UTILIZADA

A base de dados selecionada para os Estudos de Casos foi a '*Reuters-21578, Distribution 1.0*', por ser uma base já amplamente utilizada em pesquisas relacionadas à categorização de textos, e, portanto, consolidada para esse fim, e por atender às demandas desta pesquisa em termos de qualidade e quantidade de dados.

Essa base foi reunida com o objetivo de ser uma base de testes para tarefas de categorização de textos. Os seus direitos de autoria pertencem ao grupo '*Reuters Ltda and Carnegie Group*', que promoveu a sua livre distribuição para atividades de pesquisa.

Ela possui 21.578 registros de documentos de 120 categorias distintas, de diversas áreas do conhecimento, como Ciência, Tecnologia, Economia, Negócios, Agricultura, Petróleo e Mineração, entre outros. Esses registros estão dispostos em 22 arquivos eletrônicos (os 21 primeiros contendo 1.000 registros cada e o último contendo 578 registros). Os documentos estão no idioma Inglês.

Os textos se encontram em arquivos no formato *SGML* (*Standard Generalized Markup Language*), que originou formatos bastante difundidos atualmente, como o *HTML* e o *XML*.

### 4.1.1. PRÉ-PROCESSAMENTO GERAL DOS DADOS

Os textos, como mencionado no item 4.1, se encontram no formato *SGML*. O *SGML*, uma linguagem de marcação genérica, é um padrão internacional que define regras para o uso de marcações descritivas das seções, as *TAGs*, mescladas ao texto.

A figura 4.1 apresenta o exemplo de um texto retirado da base de dados, ainda sem nenhum processamento, que foi utilizado neste trabalho, da classe '*TRADE*'.

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="12521" NEWID="338">
<DATE> 2-MAR-1987 07:19:13.49</DATE>
<TOPICS><D>TRADE</D></TOPICS>
<PLACES><D>sweden</D><D>south-africa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;RM
&#22;&#22;&#1;f0078&#31;reuteu f BC-SWEDEN-TO-GO-AHEAD-WI   03-02
0094</UNKNOWN>
<TEXT>&#2;
<TITLE>SWEDEN TO GO AHEAD WITH S. AFRICAN TRADE SANCTIONS</TITLE>
<DATELINE> STOCKHOLM, March 2 </DATELINE>
<BODY>Sweden's ruling Social Democratic Party gave full power to the government to
decree unilateral trade sanctions against S. Africa, Prime Minister Ingvar Calrsson said.
        Carlsson told a news conference the party decided the fight against apartheid took
priority over Sweden's traditional policy of only adopting sanctions with the backing of the
U.N. Security Council.
        The government will decide later what form the trade boycott will take and when it
will come into force.
REUTER
&#3;</BODY></TEXT>
</REUTERS>
```

Figura 4.1 : Exemplo de texto original retirado da base de dados utilizada

Os textos utilizados possuem uma série de seções que não apresentam conteúdo algum, ou seja, encontram-se vazias (como, no exemplo do texto da figura 4.1, a seção

‘<PEOPLE> </PEOPLE>’). Outras seções apresentam conteúdo que nada acrescenta às tarefas realizadas (como, no exemplo do texto da figura 4.1, a seção ‘<UNKNOWN> </UNKNOWN>’).

Avaliando-se os textos, conclui-se que as duas únicas seções de interesse para essa pesquisa eram as seções ‘<TITLE> </TITLE>’ e ‘<BODY> </BODY>’.

O texto no corpo do documento (no exemplo do texto da figura 4.1 na seção ‘<BODY> </BODY>’) apresenta alguns caracteres inválidos, não reconhecidos pelo *ORACLE TEXT*, como “**&#**” e “ **'** ” , que provocavam erros durante a inserção dos dados no banco.

Foi necessário, então, um amplo pré-processamento nos textos utilizados, para o tratamento dos dados, em que foram retirados os caracteres inválidos e excluídas as seções que não seriam utilizadas.

A figura 4.2 mostra o exemplo do texto apresentado na figura 4.1 após o pré-processamento dos dados.

<TITLE>SWEDEN TO GO AHEAD WITH S. AFRICAN TRADE SANCTIONS</TITLE>
<BODY> Sweden´s ruling Social Democratic Party gave full power to the government to decree unilateral trade sanctions against S. Africa, Prime Minister Ingvar Calrsson said.

Carlsson told a news conference the party decided the fight against apartheid took priority over Sweden´s traditional policy of only adopting sanctions with the backing of the U.N. Security Council.

The government will decide later what form the trade boycott will take and when it will come into force.</BODY>

Figura 4.2 : Exemplo de texto após a etapa de pré-processamento

## 4.1.2. AS CLASSES UTILIZADAS

Dentre as 120 classes existentes na base de dados procurou-se investigar as 20 mais numerosas em termos de quantidade de documentos, a fim de se ter disponível um amplo material para o trabalho. A figura 4.3 apresenta as 20 classes analisadas.

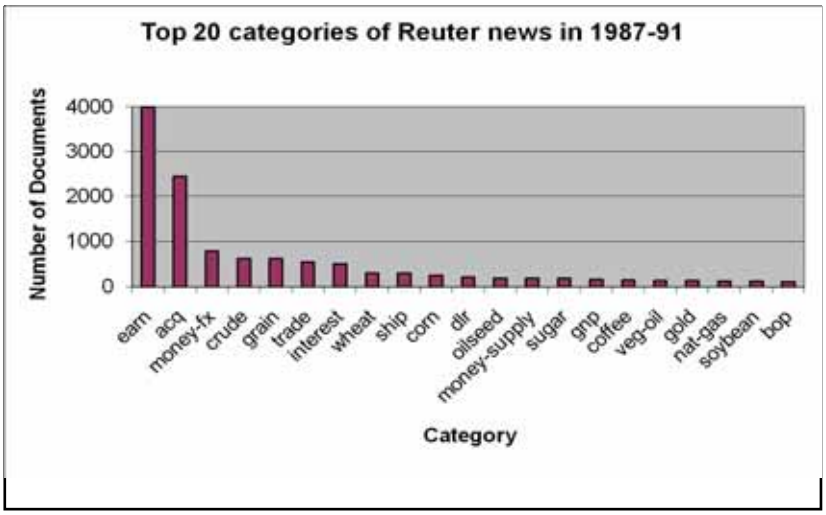


Figura 4.3 : Gráfico com as 20 classes mais numerosas da base de dados

Na base de dados existem documentos de vários tipos: documentos que pertencem apenas a uma classe, documentos que pertencem a várias classes e documentos sem nenhuma classificação.

As classes utilizadas foram selecionadas procurando-se diferenciá-las ao máximo entre si, no sentido de que quando um documento fosse de determinada classe ele não fosse ao mesmo tempo de outras classes selecionadas, a fim de não atrapalhar ou mascarar as tarefas de categorização. Na análise da base de dados, percebeu-se, por exemplo, que a maioria dos documentos que pertencia à classe '*GRAIN*' também pertencia à classe '*WHEAT*'. Daí selecionou-se somente a primeira para a realização das tarefas.

Foram eleitas 10 classes para integrar os Estudos de Casos, e um total de 120 documentos por classe. A seguir é apresentada uma breve descrição de cada uma delas:

1- **EARN** : referente a assuntos ligados a previsão de ganhos;
2- **ACQ** (Acquisitions) : referente a assuntos ligados a aquisições e fusões;
3- **CRUDE** (Oil) : referente a assuntos ligados a óleo cru (Petróleo);
4- **TRADE** : referente a assuntos ligados ao comércio e a troca;
5- **GRAIN** : referente a assuntos ligados a grãos (Agricultura);
6- **SHIP** : referente a assuntos ligados a navios e embarcações;
7- **MONEY** : referente a assuntos ligados ao dinheiro e ao câmbio externo;
8- **SUGAR** : referente a assuntos ligados ao açúcar;
9- **COFFEE** : referente a assuntos ligados ao café;
10- **GOLD** : referente a assuntos ligados ao ouro.

No item 4.3 será especificado que classes foram utilizadas em cada caso, a quantidade de documentos adotada, e os demais detalhes.

## 4.2 AS TAREFAS REALIZADAS

Esta seção apresenta as tarefas que foram realizadas nos Estudos de Casos. Elas estão agrupadas de acordo com a ferramenta utilizada, ou seja, tarefas realizadas com o *ORACLE TEXT* (*OT*) e tarefas realizadas com o *Poly Analyst* (*PA*).

### 4.2.1. TAREFAS REALIZADAS COM O *ORACLE TEXT*

Abaixo apresentamos as tarefas que foram realizadas com o *ORACLE TEXT*:

- Classificação baseada em Regras;
- Classificação Supervisionada baseada em Árvores de Decisão;
- Classificação Supervisionada baseada no algoritmo *SVM*;
- Clusterização;
- Visualização.

### 4.2.2. TAREFAS REALIZADAS COM O *POLY ANALYST*

Segue abaixo as tarefas realizadas com o *Poly Analyst*:

- Classificação Linear baseada no algoritmo Bayesiano Simples;
- Classificação Linear baseada no algoritmo *SVM*;
- Clusterização.

Não foram realizadas tarefas de visualização no *PA* pois os resultados que são produzidos não encontram equivalência com os resultados produzidos pelo *OT*, e a intenção do uso do *PA* é apenas validar os resultados obtidos com o Oracle.

## 4.3 OS ESTUDOS DE CASOS

Na composição dos Estudos de Casos, um dos principais requisitos definidos foi poder avaliar a escalabilidade do sistema, ou seja, a sua capacidade de tratar conjuntos grandes e diversos de dados. A acurácia dos resultados obtidos, a complexidade da configuração e da execução das tarefas, a apresentação dos resultados, além do uso dos recursos computacionais e do tempo de resposta do sistema foram outros fatores considerados na elaboração e na avaliação do desempenho do sistema nos EC.

Optou-se por elaborar cinco EC:

- Os Estudos de Casos **EC1**, **EC2**, **EC3** e **EC4**: para a execução das tarefas de Classificação (baseada em Regras e as Supervisionadas) e Clusterização no *OT*, com quantidades distintas de classes e de registros, para que o desempenho da ferramenta pudesse ser avaliado de forma abrangente, com a validação dos resultados a partir da comparação com os obtidos com o *PA*. A tabela 4.1 traz um resumo da configuração geral dos EC, no que diz respeito às categorias envolvidas e à quantidade de documentos em cada caso.

| ESTUDOS DE CASO<br>CATEGORIAS | EC1 | EC2 | EC3 | EC4 |
|---|---|---|---|---|
| **1- EARN** | 30 | 120 | 30 | 120 |
| **2- ACQ** | 30 | 120 | 30 | 120 |
| **3- CRUDE** | 30 | 120 | 30 | 120 |
| **4- TRADE** | 30 | 120 | 30 | 120 |
| **5- GRAIN** | 30 | 120 | 30 | 120 |
| **6- SHIP** | 30 | 120 | 30 | 120 |
| **7- MONEY** | | | 30 | 120 |
| **8- SUGAR** | | | 30 | 120 |
| **9- COFFEE** | | | 30 | 120 |
| **10- GOLD** | | | 30 | 120 |
| **TOTAL DE DOCUMENTOS** | **180** | **720** | **300** | **1200** |

Tabela 4.1: Tabela resumo com a configuração dos Estudos de Caso

- O Estudo de Caso **EC5**: para a execução e a avaliação das tarefas de Visualização dispostas no *OT*.

Para as tarefas de Classificação Supervisionada no *OT*, foram separados 2/3 dos documentos para a base de treinamento. O 1/3 restante foi utilizado para a Classificação propriamente dita.

Para se ter a mesma base de avaliação, na Classificação Baseada em Regras foi utilizado o mesmo 1/3 dos documentos disponíveis (embora fosse possível usar a totalidade deles, já que, nesse caso, não era necessário dotar o sistema de um conjunto de treinamento e sim de uma tabela com a definição das categorias e com a descrição das regras).

Para efeito de comparação e validação, nas tarefas de Classificação do *PA* foi utilizado também o mesmo 1/3 dos documentos usados no *OT*.

Nas tarefas de Classificação baseada em Regras no *OT* é preciso fornecer ao sistema uma tabela com a descrição das categorias e as regras de definição de cada uma, conforme mencionado anteriormente. As regras que definem as categorias compreendem, basicamente, a enumeração de termos a serem buscados nos documentos.

A tabela 4.2 foi desenvolvida e utilizada no sistema com essa finalidade. Ela mostra as regras desenvolvidas para cada categoria.

| ID | CATEGORIA | REGRAS |
|---|---|---|
| **1** | **EARN** | Earn or Shareholder or Share or Gain or Mln Or Million or Dlr or Dollar or Stockholder or Dividend or Pay or Capital |
| **2** | **ACQ** | Acquisition or Merger or Purchase or Acquire or Acquired or Company or Investment or Stock |
| **3** | **CRUDE** | Crude or Oil or Barrel or Reserve or Petroleum or Refinery or Exploration or Prospect or Opec |
| **4** | **TRADE** | Trade or Business or Commerce or Economy or Price or Export or Exports or Import or Tariff or Economic or Credit |
| **5** | **GRAIN** | Grain or Wheat or Pesticide or Agriculture or Plant or Cereal or Corn or Ton or Tonnes or Farm or Legume |
| **6** | **SHIP** | Ship or Port or Coast or Seafarer or Canal or Barge or Gulf or Lake or Maritime or Sea or Marine or Shipping or Boat or Pier |
| **7** | **MONEY** | Money or Foreign or Exchange or Yen or Bank or Deposit or Monetary or Fund or Reserve or Dollar or Peso or Euro or Real or Libra or Mln or Million |
| **8** | **SUGAR** | Sugar or Sugarcane or Cane or Ton or Tonnes or Crystal |
| **9** | **COFFEE** | Coffee or Ico or Crop or Tea |
| **10** | **GOLD** | Gold or Mine or Miner or Mining or Precious or Metal or Platinum or Silver or Exploration |

Tabela 4.2: Tabela com a descrição e as regras de cada categoria

Na apresentação dos EC, nas tabelas com os resultados, os valores que aparecem em preto dizem respeito aos registros que foram corretamente classificados, enquanto que os que aparecem em cinza dizem respeito aos que não foram.

## 4.3.1. ESTUDO DE CASO 1 (EC1)

Integram esse EC1 as seis classes listadas: 1- EARN, 2- ACQ, 3- CRUDE, 4- TRADE, 5- GRAIN e 6- SHIP.

Para compor a base de dados foram selecionados 30 documentos de cada uma dessas 6 classes, totalizando 180 documentos.

Para as tarefas de Classificação Supervisionada no *OT*, foram separados 2/3 dos documentos para a base de treinamento, ou seja, 20 documentos de cada classe, totalizando 120 documentos. Daí, o 1/3 restante, com 10 documentos por classe, totalizando 60 documentos, foi utilizado para a Classificação propriamente dita.

Esse é o menor EC elaborado neste trabalho.

### 4.3.1.1. EC1 - *ORACLE TEXT* (*OT*)

EC1 - *OT* - CLASSIFICAÇÃO BASEADA EM REGRAS

As regras de definição das classes utilizadas estão dispostas na Tabela 4.1 apresentada no item 4.3 .

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **10** | | | | | | 10 | 100 |
| 10 | **ACQ** | | **10** | | | | | 10 | 100 |
| 10 | **CRUDE** | | | **10** | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | 10 | 100 |
| **60** | **TOTAL** | **10** | **10** | **10** | **10** | **10** | **10** | **60** | **100** |

Tabela 4.3: EC1 – *OT* – Resultados da Classificação baseada em Regras

EC1 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ÁRVORES DE DECISÃO

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **10** | | | | | | 10 | 100 |
| 10 | **ACQ** | 1 | **9** | | | | | 10 | 90 |
| 10 | **CRUDE** | | | **10** | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | 10 | 100 |
| **60** | **TOTAL** | **11** | **9** | **10** | **10** | **10** | **10** | **60** | **98,3** |

Tabela 4.4: EC1 – *OT* – Resultados da Classificação Supervisionada – AD

EC1 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **10** | | | | | | 10 | 100 |
| 10 | **ACQ** | | **10** | | | | | 10 | 100 |
| 10 | **CRUDE** | | | **10** | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | 10 | 100 |
| **60** | **TOTAL** | **10** | **10** | **10** | **10** | **10** | **10** | **60** | **100** |

Tabela 4.5: EC1 – *OT* – Resultados da Classificação Supervisionada – SVM

EC1 - *OT* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

É possível observar que as três metodologias de Classificação obtiveram excelentes resultados em termos de precisão para esse EC1. Apenas 1 documento da classe ‘ACQ’ não foi classificado corretamente na abordagem de Classificação Supervisionada baseada em Árvores de Decisão. As regras de Classificação geradas manualmente para o método de Classificação baseada em Regras se mostraram altamente eficazes para esse conjunto de dados.

EC1 - *OT* - CLUSTERIZAÇÃO

Como o número de grupos é conhecido, o parâmetro ‘K’ do algoritmo *K-Means* utilizado pela *procedure* de Clusterização recebeu o valor 6.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | **EARN** | **20** | 1 | | 9 | | | 30 | 66,7 | Shareholder, Resource, Payment, Dividend |
| 30 | **ACQ** | | **16** | | 14 | | | 30 | 53,3 | Acquisition, Corporations, Merger, Values |
| 30 | **CRUDE** | | | 10 | **18** | | 2 | 30 | 33,3 | Oil, Organization, Finance, Production |
| 30 | **TRADE** | | | | **29** | | 1 | 30 | 96,7 | Trade, Commerce, Export, Economy |
| 30 | **GRAIN** | | | | 5 | **25** | | 30 | 83,3 | Grain, Wheat, Agriculture, Farmers |
| 30 | **SHIP** | | | | **27** | | 3 | 30 | 10 | Ships, Negotiations, Tariff, Market |
| **180** | **TOTAL** | **20** | **17** | **10** | **102** | **25** | **6** | **180** | **57,2** | |

Tabela 4.6: EC1 – *OT* – Resultados da Clusterização

O resultado geral da Clusterização realizada pode ser observado na última linha da tabela 4.6, contendo os totais. Esse seria o resultado final caso a base de dados não fosse supervisionada e, para a sua validação, seria necessário à análise de um especialista. Como essa base é supervisionada, foi possível fazer o papel do especialista e alocar os documentos aos grupos originais, além de associar os grupos gerados a cada um dos grupos inicialmente conhecidos. Com isso, foi possível fazer a avaliação da acurácia da tarefa, conforme o apresentado na tabela 4.6.

É possível observar que, em linhas gerais, o método teve um desempenho mediano em termos de precisão, com 57,2% de acertos na média. Os grupos ‘TRADE’ e ‘GRAIN’ obtiveram um desempenho muito bom, sobretudo o primeiro, enquanto os grupos ‘EARN’ e ‘ACQ’ tiveram um desempenho apenas médio. O grupo ‘CRUDE’ teve um desempenho não satisfatório, enquanto que o grupo ‘SHIP’ teve um desempenho muito ruim, com apenas 10% de acertos. O grupo ‘TRADE’ encontrado concentra 102 dos 180 documentos (56,7% do total).

Vale assinalar que a técnica abordou 100% dos documentos envolvidos.

### 4.3.1.2. EC1 - *POLY ANALYST* (*PA*)

EC1 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO BAYESIANO SIMPLES

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **9** | 1 | | | | | 10 | 90 |
| 10 | **ACQ** | | **10** | | | | | 10 | 100 |
| 10 | **CRUDE** | | | **10** | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | 10 | 100 |
| **60** | **TOTAL** | **9** | **11** | **10** | **10** | **10** | **10** | **60** | **98,3** |

Tabela 4.7: EC1 – *PA* – Resultados da Classificação Linear – BS

EC1 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **10** | | | | | | 10 | 100 |
| 10 | **ACQ** | | **10** | | | | | 10 | 100 |
| 10 | **CRUDE** | | | **10** | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | 10 | 100 |
| **60** | **TOTAL** | **10** | **10** | **10** | **10** | **10** | **10** | **60** | **100** |

Tabela 4.8: EC1 – *PA* – Resultados da Classificação Linear – *SVM*

EC1 - *PA* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

É possível notar que as duas abordagens de Classificação obtiveram resultados muito bons em termos de precisão, com apenas 1 documento não classificado corretamente na abordagem de Classificação baseada no algoritmo Bayesiano Simples. Isso demonstra que as abordagens dispostas no programa são adequadas para o tratamento desse conjunto de dados e que nele as categorias se encontram

suficientemente separadas espacialmente, uma vez que diferentes métodos de Classificação obtiveram resultados excelentes em termo de precisão.

EC1 - *PA* - CLUSTERIZAÇÃO

Depois de alguns experimentos, foi adotada uma configuração de parâmetros que resultou na descoberta de 6 grupos.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|
| 30 | **EARN** | **14** | | | | | | 14 | 46,7 | Cts, qtr, shr |
| 30 | **ACQ** | 4 | **14** | | | 1 | 2 | 21 | 46,7 | Inc, Share |
| 30 | **CRUDE** | | | **23** | | 4 | 1 | 28 | 76,7 | Oil, Price, Petroleum |
| 30 | **TRADE** | | | 2 | **26** | 12 | 2 | 42 | 86,7 | Trade, Export |
| 30 | **GRAIN** | **12** | 11 | 5 | 3 | 9 | 4 | 44 | 30 | Dlr, mln |
| 30 | **SHIP** | | 1 | | 1 | | **10** | 12 | 33,3 | De, Union |
| **180** | **TOTAL** | **30** | **26** | **30** | **30** | **26** | **19** | **161** | **53,4** | |

Tabela 4.9: EC1 – *PA* – Resultados da Clusterização

É possível perceber que a técnica teve um desempenho médio em termos de acurácia dos resultados, com 53,4% de acertos na média. Os grupos ‘TRADE’ e ‘CRUDE’ obtiveram bons resultados, enquanto que os grupos ‘EARN’ e ‘ACQ’ obtiveram resultados médios em termos de precisão. Já os grupos ‘GRAIN’ e ‘SHIP’ obtiveram resultados considerados insatisfatórios, na faixa dos 30% de acertos.

Apesar de a média dos resultados alcançados aqui ser próxima da média obtida com o *OT*, numa avaliação mais detalhada são observadas diferenças significativas. Isso, de certa forma, não foi nenhuma surpresa, uma vez que as abordagens utilizam técnicas distintas para essa tarefa: enquanto o *OT* utiliza o algoritmo *K-MEANS* para a descoberta dos grupos, o *PA* utiliza um algoritmo que é uma variação do algoritmo *Suffix Tree Clustering*, que, entre outras coisas, não requer uma definição do número de grupos a ser encontrado.

Vale assinalar que, diferentemente do observado com o *OT*, o método abordou apenas 161 dos 180 documentos envolvidos, o que corresponde a 89,4% do total.

## 4.3.2. ESTUDO DE CASO 2 (EC2)

Integram esse EC2 as seis classes listadas: 1- EARN, 2- ACQ, 3- CRUDE, 4- TRADE, 5- GRAIN e 6- SHIP.

Para compor a base de dados foram selecionados 120 documentos de cada uma dessas 6 classes, totalizando 720 documentos.

Para as tarefas de Classificação Supervisionada, foram separados 2/3 dos documentos para a base de treinamento, ou seja, 80 documentos de cada classe, totalizando 480 documentos. Daí, o 1/3 restante, com 40 documentos por classe, totalizando 240 documentos, foi utilizado para a Classificação propriamente dita.

### 4.3.2.1. EC2 - *ORACLE TEXT* (*OT*)

EC2 - *OT* - CLASSIFICAÇÃO BASEADA EM REGRAS

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **39** | | | | | | 39 | 97,5 |
| 40 | **ACQ** | 1 | **39** | | | | | 40 | 97,5 |
| 40 | **CRUDE** | | | **40** | | | | 40 | 100 |
| 40 | **TRADE** | 1 | | | **39** | | | 40 | 97,5 |
| 40 | **GRAIN** | | | | 1 | **39** | | 40 | 97,5 |
| 40 | **SHIP** | | 1 | 1 | | | **38** | 40 | 95 |
| **240** | **TOTAL** | **41** | **40** | **41** | **40** | **39** | **38** | **239** | **97,5** |

Tabela 4.10: EC2 – *OT* – Resultados da Classificação baseada em Regras

EC2 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ÁRVORES DE DECISÃO

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **40** | | | | | | 40 | 100 |
| 40 | **ACQ** | 1 | **38** | | 1 | | | 40 | 95 |
| 40 | **CRUDE** | | | **39** | 1 | | | 40 | 97,5 |
| 40 | **TRADE** | 1 | 1 | | **38** | | | 40 | 95 |
| 40 | **GRAIN** | | | | 1 | **39** | | 40 | 97,5 |
| 40 | **SHIP** | | | | | | **40** | 40 | 100 |
| **240** | **TOTAL** | **42** | **39** | **39** | **41** | **39** | **40** | **240** | **97,5** |

Tabela 4.11: EC2 – *OT* – Resultados da Classificação Supervisionada – AD

EC2 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **40** | | | | | | 40 | 100 |
| 40 | **ACQ** | 1 | **39** | | | | | 40 | 97,5 |
| 40 | **CRUDE** | | | **40** | | | | 40 | 100 |
| 40 | **TRADE** | 1 | | | **39** | | | 40 | 97,5 |
| 40 | **GRAIN** | | | | | **40** | | 40 | 100 |
| 40 | **SHIP** | | | | | | **40** | 40 | 100 |
| **240** | **TOTAL** | **42** | **39** | **40** | **39** | **40** | **40** | **240** | **99,2** |

Tabela 4.12: EC2 – *OT* – Resultados da Classificação Supervisionada – SVM

EC2 - *OT* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

Analisando os dados das tabelas 4.10 a 4.12 é possível observar que tanto a abordagem de Classificação baseada em Regras quanto às abordagens de Classificação Supervisionada obtiveram resultados muito bons com relação à acurácia, com um pequeno destaque para a Classificação Supervisionada baseada no *SVM*, com 99,2% de acertos na média. Na Classificação baseada em Regras um documento da classe 'EARN' não foi classificado.

Esse Estudo de Caso é considerado de porte médio e é quatro vezes maior que o EC1 e, mesmo assim, o *OT* obteve resultados equivalentes aos daquele EC, com uma margem muito pequena de diferença.

EC2 - *OT* - CLUSTERIZAÇÃO

O parâmetro ‘K’ do algoritmo *K-Means* utilizado pela *procedure* de Clusterização recebeu o valor 6.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|
| 120 | **EARN** | **74** | 36 | | 10 | | | 120 | 61,7 | Shareholder, Resource, Payment, Dividend |
| 120 | **ACQ** | | **118** | | 2 | | | 120 | 98,3 | Acquisition, Corporations, Merger, Values |
| 120 | **CRUDE** | | **70** | 42 | 8 | | | 120 | 35 | Oil, Organization, Finance, Production |
| 120 | **TRADE** | | 56 | | **64** | | | 120 | 53,3 | Trade, Commerce, Export, Economy |
| 120 | **GRAIN** | | 18 | | 2 | **100** | | 120 | 83 | Grain, Wheat, Agriculture, Farmers |
| 120 | **SHIP** | | **90** | | 2 | | 28 | 120 | 23,3 | Ships, Negotiations, Tariff, Market |
| **720** | **TOTAL** | **74** | **388** | **42** | **88** | **100** | **28** | **720** | **59,1** | |

Tabela 4.13: EC2 – *OT* – Resultados da Clusterização

Analisando-se a tabela 4.13 é possível observar que a abordagem obteve um desempenho mediano em termos de precisão, com média de 59,1% de acertos. Os grupos ‘ACQ’ e ‘GRAIN’ tiveram um desempenho muito bom, com destaque para o primeiro, enquanto os grupos ‘EARN’ e ‘TRADE’ obtiveram desempenhos médios. Os grupos ‘CRUDE’ e ‘SHIP’ tiveram desempenhos considerados não satisfatórios. O grupo ‘ACQ’ encontrado concentra 388 dos 720 documentos (53,9% do total).

Há uma variação significativa com relação ao EC1, onde o grupo ‘TRADE’ teve o melhor resultado, com quase 100% de acerto, e o grupo ‘ACQ’ obteve um resultado próximo dos 50% de acertos. Isso pode ser explicado pela proximidade que acredita-se existir entre esses dois grupos.

Cabe destacar que neste EC2 o *OT* teve um desempenho ligeiramente superior ao obtido no EC1 nessa tarefa, mesmo este tendo quatro vezes o tamanho do primeiro, o que pode ser associado às características dos documentos desta base de dados.

Esse método abordou 100% dos documentos envolvidos.

### 4.3.2.2. EC2 - *POLY ANALYST* (*PA*)

EC2 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO BAYESIANO SIMPLES

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **39** | 1 | | | | | 40 | 100 |
| 40 | **ACQ** | 2 | **37** | | 1 | | | 40 | 95 |
| 40 | **CRUDE** | | | **39** | 1 | | | 40 | 97,5 |
| 40 | **TRADE** | | 1 | | **39** | | | 40 | 100 |
| 40 | **GRAIN** | | | | | **40** | | 40 | 100 |
| 40 | **SHIP** | | | | | | **40** | 40 | 100 |
| **240** | **TOTAL** | **41** | **39** | **39** | **41** | **40** | **40** | **240** | **97,5** |

Tabela 4.14: EC2 – *PA* – Resultados da Classificação Linear – BS

EC2 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **40** | | | | | | 40 | 100 |
| 40 | **ACQ** | 1 | **39** | | | | | 40 | 97,5 |
| 40 | **CRUDE** | | | **40** | | | | 40 | 100 |
| 40 | **TRADE** | | | | **40** | | | 40 | 100 |
| 40 | **GRAIN** | | | | | **40** | | 40 | 100 |
| 40 | **SHIP** | | | | | | **40** | 40 | 100 |
| **240** | **TOTAL** | **41** | **39** | **40** | **40** | **40** | **40** | **240** | **99,6** |

Tabela 4.15: EC2 – *PA* – Resultados da Classificação Linear – SVM

EC2 - *PA* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

É possível notar nas tabelas 4.14 e 4.15 que os dois métodos de Classificação apresentados alcançaram excelentes resultados com relação à precisão, com destaque para a Classificação baseada no *SVM*, com 99,6% de acertos. Vale citar que esses resultados são muito equivalentes aos alcançados com o *OT* para esse EC, o que indica que as abordagens dispostas no *OT* são adequadas para tratar esse conjunto de documentos.

EC2 - *PA* - CLUSTERIZAÇÃO

Após alguns testes iniciais, foi adotada uma configuração de parâmetros que resultou na descoberta de 6 grupos.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|
| 120 | **EARN** | **100** | 22 | 22 | 16 | 10 | 2 | 172 | 83,3 | Cts, Net, shr, Note |
| 120 | **ACQ** | | 2 | **12** | | | 6 | 20 | 1,7 | Gasolene, Combustible |
| 120 | **CRUDE** | 2 | | **82** | 2 | 2 | | 88 | 68,3 | Crude, Barrel |
| 120 | **TRADE** | | | | 10 | 2 | **40** | 52 | 8,3 | Gob, Leader, Janeiro |
| 120 | **GRAIN** | | | 2 | 4 | **104** | 16 | 126 | 86,7 | Mt, Wheat |
| 120 | **SHIP** | | | 6 | 2 | 2 | **64** | 74 | 53 | Port, Strike |
| **720** | **TOTAL** | **102** | **24** | **124** | **34** | **120** | **128** | **532** | **50,2** | |

Tabela 4.16: EC2 – *PA* – Resultados da Clusterização

Podemos observar na tabela 4.16 que esse método obteve um desempenho mediano com relação à precisão dos resultados, com 50,2% de acertos na média. Os grupos ‘EARN’ e ‘GRAIN’ obtiveram bons resultados, na faixa dos 85% de acertos, enquanto que os grupos ‘CRUDE’ e ‘SHIP’ obtiveram resultados apenas médios. Já os grupos ‘TRADE’ e ‘ACQ’ obtiveram resultados muito ruins, abaixo dos 10% de acertos, sendo que o último conseguiu agrupar corretamente menos de 2% dos documentos.

Apesar de a observação detalhada dos grupos gerados aqui denotar grandes diferenças em relação aos grupos encontrados com o *OT*, a % de acertos na média foi muito próxima entre os dois programas, o que talvez possa ter como causa as características gerais dessa base de dados, ainda que as duas ferramentas tenham abordagens distintas para a realização dessa tarefa.

Essa técnica tratou somente de 532 dos 720 documentos, o que corresponde a 73,9% do total.

## 4.3.3. ESTUDO DE CASO 3 (EC3)

Integram esse EC3 as dez classes listadas: 1- EARN, 2- ACQ, 3- CRUDE, 4- TRADE, 5- GRAIN, 6- SHIP, 7- MONEY, 8- SUGAR, 9- COFFEE e 10- GOLD.

Para compor a base de dados foram selecionados 30 documentos de cada uma dessas 10 classes, totalizando 300 documentos.

Para as tarefas de Classificação Supervisionada, foram separados 2/3 dos documentos para a base de treinamento, ou seja, 20 documentos de cada classe, totalizando 200 documentos. Daí, o 1/3 restante, com 10 documentos por classe, totalizando 100 documentos, foi utilizado para a Classificação propriamente dita.

### 4.3.3.1. EC3 - *ORACLE TEXT* (*OT*)

EC3 - *OT* - CLASSIFICAÇÃO BASEADA EM REGRAS

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **10** | | | | | | | | | | 10 | 100 |
| 10 | **ACQ** | | **10** | | | | | | | | | 10 | 100 |
| 10 | **CRUDE** | | | **10** | | | | | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | | | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | | | | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | | | | | 10 | 100 |
| 10 | **MONEY** | | | | | | | **10** | | | | 10 | 100 |
| 10 | **SUGAR** | | | | | | | | **10** | | | 10 | 100 |
| 10 | **COFFEE** | | | | | | | | | **10** | | 10 | 100 |
| 10 | **GOLD** | | | | | | | | | | **10** | 10 | 100 |
| **100** | **TOTAL** | **10** | **10** | **10** | **10** | **10** | **10** | **10** | **10** | **10** | **10** | **100** | **100** |

Tabela 4.17: EC3 – *OT* – Resultados da Classificação baseada em Regras

EC3 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ÁRVORES DE DECISÃO

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **10** | | | | | | | | | | 10 | 100 |
| 10 | **ACQ** | 1 | **9** | | | | | | | | | 10 | 90 |
| 10 | **CRUDE** | | | **10** | | | | | | | | 10 | 100 |
| 10 | **TRADE** | | | | **9** | | | 1 | | | | 10 | 90 |
| 10 | **GRAIN** | | | | | **10** | | | | | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | | | | | 10 | 100 |
| 10 | **MONEY** | 1 | | | | | | **9** | | | | 10 | 90 |
| 10 | **SUGAR** | | | | | | | | **10** | | | 10 | 100 |
| 10 | **COFFEE** | | | | | 1 | | | | **9** | | 10 | 90 |
| 10 | **GOLD** | | | | | | | | | | **10** | 10 | 100 |
| **100** | **TOTAL** | **12** | **9** | **10** | **9** | **11** | **10** | **10** | **10** | **9** | **10** | **100** | **96** |

Tabela 4.18: EC3 – *OT* – Resultados da Classificação Supervisionada – AD

EC3 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **10** | | | | | | | | | | 10 | 100 |
| 10 | **ACQ** | | **9** | | | | | 1 | | | | 10 | 90 |
| 10 | **CRUDE** | | | **10** | | | | | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | | | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | | | | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | | | | | 10 | 100 |
| 10 | **MONEY** | | | | 1 | | | **9** | | | | 10 | 90 |
| 10 | **SUGAR** | | | | | | | | **10** | | | 10 | 100 |
| 10 | **COFFEE** | | | | | | | | | **10** | | 10 | 100 |
| 10 | **GOLD** | | | | | | | | | | **10** | 10 | 100 |
| **100** | **TOTAL** | **10** | **9** | **10** | **11** | **10** | **10** | **10** | **10** | **10** | **10** | **100** | **98** |

Tabela 4.19: EC3 – *OT* – Resultados da Classificação Supervisionada – *SVM*

EC3 - *OT* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

Com a análise das tabelas 4.17 a 4.19 é possível observar que as três abordagens de Classificação obtiveram excelentes resultados em termos de precisão. O destaque vai para a Classificação baseada em Regras, que obteve 100% de acertos. Isso confirma a

qualidade das regras geradas manualmente para essa tarefa, apresentadas na Tabela 4.2 do item 4.3. Nas tarefas de Classificação Supervisionada, as regras geradas automaticamente nas duas abordagens também se mostraram com alta capacidade de predição, conforme atestado pelos resultados obtidos.

Esse EC3 é considerado de pequeno porte para as tarefas de Classificação, com um total de 100 documentos disponíveis, embora seja mais complexo que o EC1, também de pequeno porte, já que apresenta 10 categorias distintas. Os resultados obtidos apontam que as abordagens dispostas no *OT* para tarefas de Classificação são adequadas para o tratamento desse conjunto de dados.

EC3 - *OT* - CLUSTERIZAÇÃO

O parâmetro 'K' do algoritmo *K-Means* utilizado pela *procedure* de Clusterização recebeu o valor 10.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30 | **EARN** | **21** | 9 | | | | | | | | | 30 | 70 | Shareholder,Payment Resource, Dividend |
| 30 | **ACQ** | | **29** | | | | | 1 | | | | 30 | 96,7 | Acquisition, Values, Corporations, Merger |
| 30 | **CRUDE** | | **16** | 11 | | | 2 | | | 1 | | 30 | 36,7 | Oil, Organization, Finance, Production |
| 30 | **TRADE** | | **22** | | 5 | | 2 | 1 | | | | 30 | 16,7 | Trade, Commerce, Export, Economy |
| 30 | **GRAIN** | | 3 | | 2 | **20** | | | 5 | | | 30 | 66,7 | Grain, Wheat, Agriculture, Farmers |
| 30 | **SHIP** | | **25** | | 1 | 1 | 2 | | | 1 | | 30 | 6,7 | Ships, Negotiations, Tariff, Market |
| 30 | **MONEY** | | **17** | | 1 | | | 12 | | | | 30 | 40 | Money, Balances, Credits, Capital |
| 30 | **SUGAR** | | | | | 1 | | | **29** | | | 30 | 96,7 | Sugar, Calories, Cane, Agricultural |
| 30 | **COFFEE** | | | | | | | | | **30** | | 30 | 100 | Coffee, Production, Plans, Consumer |
| 30 | **GOLD** | | 2 | | | | | | | | **28** | 30 | 93,3 | Gold, Metals, Mines, Exploration |
| **300** | **TOTAL** | **21** | **123** | **11** | **9** | **22** | **6** | **14** | **34** | **32** | **28** | **300** | **62,3** | |

Tabela 4.20: EC3 – *OT* – Resultados da Clusterização

É possível observar na tabela 4.20 que, em linhas gerais, a abordagem obteve um desempenho acima da média em termos de precisão, com 62,3% de acertos no total.

O grupo ‘ACQ’ concentrou 123 dos 300 documentos (41% do total), tendo a maioria dos documentos do seu próprio grupo e dos grupos ‘CRUDE’, ‘TRADE’, ‘SHIP’ e ‘MONEY’. Acredita-se que isso possa ser explicado em parte pelas características gerais dos dados e em parte pelo área de interseção que esses grupos parecem apresentar com o grupo ‘ACQ’. Os grupos ‘SUGAR’, ‘COFFEE’ e ‘GOLD’ apresentaram excelentes resultados, o que pode ser creditado à especificidade deles, com destaque para o grupo ‘COFFEE’, com 100% de acertos.

Vale assinalar que nesse EC3, com um total de 300 documentos em 10 grupos, o programa teve um desempenho superior ao EC1, que apresenta 180 documentos em 6 grupos, com 5,1% de acertos a mais que o EC1, o que talvez possa ser creditado às características gerais dos documentos.

Esse método tratou de 100% dos documentos envolvidos.

### 4.3.3.2. EC3 - *POLY ANALYST* (*PA*)

EC3 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO BAYESIANO SIMPLES

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **9** | 1 | | | | | | | | | 10 | 90 |
| 10 | **ACQ** | | **10** | | | | | | | | | 10 | 100 |
| 10 | **CRUDE** | | | **10** | | | | | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | | | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | | | | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | | | | | 10 | 90 |
| 10 | **MONEY** | | | | 1 | | | **9** | | | | 10 | 90 |
| 10 | **SUGAR** | | | | | | | | **10** | | | 10 | 100 |
| 10 | **COFFEE** | | | | | 1 | | | | **9** | | 10 | 100 |
| 10 | **GOLD** | | | | | | | | | | **10** | 10 | 100 |
| **100** | **TOTAL** | **9** | **11** | **10** | **11** | **11** | **10** | **9** | **10** | **9** | **10** | **100** | **97** |

Tabela 4.21: EC3 – *PA* – Resultados da Classificação Linear – BA

EC3 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | **EARN** | **9** | 1 | | | | | | | | | 10 | 90 |
| 10 | **ACQ** | | **10** | | | | | | | | | 10 | 100 |
| 10 | **CRUDE** | | | **10** | | | | | | | | 10 | 100 |
| 10 | **TRADE** | | | | **10** | | | | | | | 10 | 100 |
| 10 | **GRAIN** | | | | | **10** | | | | | | 10 | 100 |
| 10 | **SHIP** | | | | | | **10** | | | | | 10 | 100 |
| 10 | **MONEY** | | | | 1 | | | **9** | | | | 10 | 100 |
| 10 | **SUGAR** | | | | | | | | **10** | | | 10 | 100 |
| 10 | **COFFEE** | | | | | | | | | **10** | | 10 | 100 |
| 10 | **GOLD** | | | | | | | | | | **10** | 10 | 100 |
| **100** | **TOTAL** | **9** | **11** | **10** | **11** | **10** | **10** | **9** | **10** | **10** | **10** | **100** | **98** |

Tabela 4.22: EC3 – *PA* – Resultados da Classificação Linear – *SVM*

EC3 - *PA* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

É possível notar nas tabelas 4.21 e 4.22 que as duas técnicas de Classificação obtiveram resultados muito bons em termos de precisão, com leve destaque para a abordagem através do algoritmo *SVM*, o que pode ser considerado dentro da expectativa.

Os resultados encontrados com essa ferramenta são equivalentes aos encontrados com o *OT*, o que, neste trabalho, vale como validação dos resultados obtidos com o Oracle.

Nas duas abordagens de Classificação dispostas aqui 100% dos documentos foram tratados.

EC3 - *PA* - CLUSTERIZAÇÃO

Depois de alguns experimentos, foi adotada uma configuração de parâmetros que resultou na descoberta de 10 grupos.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30 | **EARN** | **25** | 1 | | | | | | | | | 26 | 83,3 | Cts, qtr, shr |
| 30 | **ACQ** | | 3 | 2 | 1 | 1 | 4 | | 1 | **11** | | 23 | 10 | Talk, Quota, Organization |
| 30 | **CRUDE** | 1 | 5 | **16** | 6 | 9 | | 4 | 1 | 2 | 4 | 48 | 53,3 | Oil, price, market |
| 30 | **TRADE** | | | | **4** | 2 | | | 2 | | | 8 | 13,3 | Administration, Economic |
| 30 | **GRAIN** | | | 4 | 1 | 3 | | | **9** | 2 | 5 | 24 | 10 | Production, Output |
| 30 | **SHIP** | 1 | 3 | 1 | 7 | | **9** | 5 | 1 | 4 | 1 | 32 | 30 | Reuter, Tell |
| 30 | **MONEY** | | | | 1 | | 3 | **16** | | | | 20 | 53,3 | Money, Market, Shortage |
| 30 | **SUGAR** | | | | | 1 | | | **12** | | | 13 | 40 | White Sugar, Sargainer |
| 30 | **COFFEE** | | | 1 | | | | | | **11** | | 12 | 36,7 | Coffee, Producer, Export |
| 30 | **GOLD** | | | | | | | | | | **13** | 13 | 43,3 | Gold, Ton, Ounce |
| **300** | **TOTAL** | **27** | **12** | **24** | **20** | **16** | **16** | **25** | **26** | **30** | **23** | **219** | **39** | |

Tabela 4.23: EC3 – *PA* – Resultados da Clusterização

Como observado na tabela 4.23, essa abordagem teve um desempenho abaixo da média em termos de precisão dos resultados, com média de 39% de acertos no total. Apenas o grupo ‘ EARN’ obteve resultados significativos, com mais de 80% de acertos. Mesmo no caso dos grupos ‘SUGAR’, ‘COFFEE’ e ‘GOLD’, considerados suficientemente específicos, onde o *OT* obteve resultados excelentes, a abordagem do *PA* não alcançou nem 50% de acertos.

Como mencionado no item 4.3.1.2, as técnicas adotadas pelo *PA* são diferentes das adotadas pelo *OT* para tarefas de Clusterização, e a abordagem disposta aqui se mostra inapropriada para tratar esse conjunto de dados.

## 4.3.4. ESTUDO DE CASO 4 (EC4)

Integram esse EC4 as dez classe listadas: 1- EARN, 2- ACQ, 3- CRUDE, 4- TRADE, 5- GRAIN, 6- SHIP, 7- MONEY, 8- SUGAR, 9- COFFEE e 10- GOLD.

Para compor a base de dados foram selecionados 120 documentos de cada uma dessas 10 classes, totalizando 1200 documentos.

Para as tarefas de Classificação Supervisionada, foram separados 2/3 dos documentos para a base de treinamento, ou seja, 80 documentos de cada classe, totalizando 800 documentos. Daí, o 1/3 restante, com 40 documentos por classe, totalizando 400 documentos, foi utilizado para a Classificação propriamente dita.

Esse é o maior EC realizado neste trabalho.

### 4.3.4.1. EC4 - *ORACLE TEXT* (*OT*)

EC4 - *OT* - CLASSIFICAÇÃO BASEADA EM REGRAS

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **39** | | | | | | | | | | 39 | 97,5 |
| 40 | **ACQ** | | **39** | | | | | 1 | | | | 40 | 97,5 |
| 40 | **CRUDE** | | | **40** | | | | | | | | 40 | 100 |
| 40 | **TRADE** | 1 | | | **39** | | | | | | | 40 | 97,5 |
| 40 | **GRAIN** | | | | | **39** | | 1 | | | | 40 | 97,5 |
| 40 | **SHIP** | | | 1 | | | **38** | 1 | | | | 40 | 95 |
| 40 | **MONEY** | | | | | | | **40** | | | | 40 | 100 |
| 40 | **SUGAR** | | | | | | | | **40** | | | 40 | 100 |
| 40 | **COFFEE** | | | | | | | | | **40** | | 40 | 100 |
| 40 | **GOLD** | | | | | | | | | | **40** | 40 | 100 |
| **400** | **TOTAL** | **40** | **39** | **41** | **39** | **39** | **38** | **43** | **40** | **40** | **40** | **399** | **98,5** |

Tabela 4.24: EC4 – *OT* – Resultados da Classificação baseada em Regras

EC4 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ÁRVORES DE DECISÃO

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **40** | | | | | | | | | | 40 | 100 |
| 40 | **ACQ** | 2 | **36** | | 1 | | | 1 | | | | 40 | 90 |
| 40 | **CRUDE** | | | **38** | 1 | | 1 | | | | | 40 | 95 |
| 40 | **TRADE** | | 1 | | **37** | | | 2 | | | | 40 | 92,5 |
| 40 | **GRAIN** | | 1 | | 1 | **38** | | | | | | 40 | 95 |
| 40 | **SHIP** | | | 1 | | | **39** | | | | | 40 | 97,5 |
| 40 | **MONEY** | 1 | | | 2 | | | **37** | | | | 40 | 92,5 |
| 40 | **SUGAR** | | | | | | | | **40** | | | 40 | 100 |
| 40 | **COFFEE** | | | | | 2 | | | | **38** | | 40 | 95 |
| 40 | **GOLD** | | | | | | | | | | **40** | 40 | 100 |
| **400** | **TOTAL** | **43** | **38** | **39** | **42** | **40** | **40** | **40** | **40** | **38** | **40** | **400** | **95,8** |

Tabela 4.25: EC4 – *OT* – Resultados da Classificação Supervisionada – AD

EC4 - *OT* - CLASSIFICAÇÃO SUPERVISIONADA - ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **40** | | | | | | | | | | 40 | 100 |
| 40 | **ACQ** | | **37** | | 1 | | | 2 | | | | 40 | 97,5 |
| 40 | **CRUDE** | | | **38** | | | 2 | | | | | 40 | 97,5 |
| 40 | **TRADE** | 1 | | | **39** | | | | | | | 40 | 100 |
| 40 | **GRAIN** | | | | | **40** | | | | | | 40 | 97,5 |
| 40 | **SHIP** | | | 2 | | | **38** | | | | | 40 | 100 |
| 40 | **MONEY** | | | | 1 | | | **39** | | | | 40 | 97,5 |
| 40 | **SUGAR** | | | | | | | | **40** | | | 40 | 100 |
| 40 | **COFFEE** | | | | | 1 | | | | **39** | | 40 | 100 |
| 40 | **GOLD** | | | | | | | | | | **40** | 40 | 100 |
| **400** | **TOTAL** | **40** | **39** | **39** | **41** | **39** | **41** | **40** | **40** | **41** | **40** | **400** | **97,5** |

Tabela 4.26: EC4 – *OT* – Resultados da Classificação Supervisionada – SVM

EC4 - *OT* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

É possível observar nas tabelas de 4.24 a 4.26 que as três abordagens de Classificação obtiveram resultados muito bons com relação à precisão, todas apresentando mais de 95% de acertos. O destaque fica por conta da Classificação

baseada em Regras, com 98,5% de acertos, o que novamente confirma a qualidade e o alto poder de predição das regras geradas manualmente. As regras geradas automaticamente pelas abordagens de Classificação Supervisionada também se mostraram altamente eficazes.

Na abordagem da Classificação baseada em Regras, ocorreu de um documento da classe 'EARN' não ter sido classificado.

Esse é o maior EC elaborado, considerado, em termos desse trabalho, de grande porte. Com os resultados alcançados aqui, é possível avaliar que as abordagens de Classificação dispostas no *OT* são adequadas para o tratamento desse conjunto de dados.

EC4 - *OT* - CLUSTERIZAÇÃO

O parâmetro 'K' do algoritmo *K-Means* utilizado pela *procedure* de Clusterização recebeu o valor 10.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 120 | **EARN** | **78** | 25 | | 17 | | | | | | | 120 | 65 | Shareholder,Payment Resource, Dividend |
| 120 | **ACQ** | | **115** | | 2 | | | 3 | | | | 120 | 95,8 | Acquisition, Values, Corporations, Merger, |
| 120 | **CRUDE** | | **58** | 42 | 7 | | 6 | 7 | | | | 120 | 35 | Oil, Organization, Finance, Production |
| 120 | **TRADE** | | **78** | | 30 | | | 12 | | | | 120 | 25 | Trade, Commerce, Export, Economy |
| 120 | **GRAIN** | | 13 | | 14 | **84** | | | 9 | | | 120 | 70 | Grain, Wheat, Agriculture, Farmers |
| 120 | **SHIP** | | **81** | | 13 | 3 | 22 | | | 1 | | 120 | 18,3 | Ships, Negotiations, Tariff, Market |
| 120 | **MONEY** | | **65** | | 5 | | | 50 | | | | 120 | 41,7 | Money, Balances, Credits, Capital |
| 120 | **SUGAR** | | | | | 3 | | | **117** | | | 120 | 97,5 | Sugar, Calories, Cane, Agricultural |
| 120 | **COFFEE** | | | | 2 | | | | | **118** | | 120 | 98,3 | Coffee, Production, Plans, Consumer |
| 120 | **GOLD** | | 5 | | | | | | | | **115** | 120 | 95,8 | Gold, Metals, Mines, Exploration |
| **1200** | **TOTAL** | **78** | **440** | **42** | **90** | **90** | **28** | **72** | **126** | **119** | **115** | **1200** | **64,2** | |

Tabela 4.27: EC4 – *OT* – Resultados da Clusterização

É possível notar na tabela 4.27 que, em linhas gerais, a abordagem teve um desempenho acima da média em termos de precisão, com 64,2% de acertos.

De forma equivalente ao ocorrido no EC3, também aqui a classe 'ACQ' concentrou a maior parte dos documentos (440 dos 1200, que corresponde a 36,7% do total), concentrando boa parte dos documentos de outras categorias. Como já mencionado, isso talvez possa ser creditado à interseção que acredita-se existir entre os documentos dessas classes com os da classe 'ACQ' e a características próprias dessa base de dados.

Esse é o maior EC elaborado, com um total de 1200 documentos de 10 classes distintas, e foi o que apresentou o melhor resultado nessa tarefa entre os quatro EC elaborados. Isso talvez possa ser creditado às características gerais dos dados.

Essa abordagem considerou 100% dos documentos envolvidos.

### 4.3.4.2. EC4 - *POLY ANALYST* (*PA*)

EC4 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO BAYESIANO SIMPLES

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **38** | 2 | | | | | | | | | 40 | 95 |
| 40 | **ACQ** | | **40** | | | | | | | | | 40 | 100 |
| 40 | **CRUDE** | 1 | | **37** | 2 | | | | | | | 40 | 92,5 |
| 40 | **TRADE** | | | | **40** | | | | | | | 40 | 100 |
| 40 | **GRAIN** | | | | | **38** | | | 1 | 1 | | 40 | 95 |
| 40 | **SHIP** | | | | | | **40** | | | | | 40 | 97,5 |
| 40 | **MONEY** | 1 | 1 | | 2 | | | **36** | | | | 40 | 92,5 |
| 40 | **SUGAR** | | | | | 1 | | | **39** | | | 40 | 100 |
| 40 | **COFFEE** | | | | | 2 | | | | **38** | | 40 | 100 |
| 40 | **GOLD** | | | | | | | | | | **40** | 40 | 100 |
| **400** | **TOTAL** | **40** | **43** | **37** | **44** | **41** | **40** | **36** | **40** | **39** | **40** | **400** | **96,5** |

Tabela 4.28: EC4 – *PA* – Resultados da Classificação Linear – BS

EC4 - *PA* - CLASSIFICAÇÃO LINEAR – ALGORITMO *SVM*

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | **EARN** | **38** | 2 | | | | | | | | | 40 | 100 |
| 40 | **ACQ** | | **37** | | 2 | | | 1 | | | | 40 | 95 |
| 40 | **CRUDE** | | | **39** | | | 1 | | | | | 40 | 92,5 |
| 40 | **TRADE** | | | | **40** | | | | | | | 40 | 100 |
| 40 | **GRAIN** | | | | | **39** | | | | 1 | | 40 | 92,5 |
| 40 | **SHIP** | | | | | | **40** | | | | | 40 | 95 |
| 40 | **MONEY** | 1 | | | 1 | | | **38** | | | | 40 | 100 |
| 40 | **SUGAR** | | | | | | | | **40** | | | 40 | 100 |
| 40 | **COFFEE** | | | | | | | | | **40** | | 40 | 97,5 |
| 40 | **GOLD** | | | | | | | | | | **40** | 40 | 100 |
| **400** | **TOTAL** | **39** | **39** | **39** | **43** | **39** | **41** | **39** | **40** | **41** | **40** | **400** | **97,8** |

Tabela 4.29: EC4 – *PA* – Resultados da Classificação Linear – *SVM*

EC4 - *PA* – CLASSIFICAÇÃO: ANÁLISE DOS RESULTADOS

É possível observar nas tabelas 4.28 e 4.29 que os dois métodos de Classificação alcançaram resultados excelentes, com um pequeno destaque para a abordagem baseada no algoritmo *SVM*.

Os resultados obtidos com essa ferramenta são equivalentes aos encontrados com o *OT*, através do que é possível validar os resultados alcançados com o Oracle.

Nos dois casos, foram tratados 100% dos documentos disponíveis.

EC4 - *PA* - CLUSTERIZAÇÃO

Após alguns experimentos iniciais, foi adotada uma configuração de parâmetros que resultou na descoberta de 10 grupos.

| DOCs | PREDIÇÃO / REAL | EARN | ACQ | CRUDE | TRADE | GRAIN | SHIP | MONEY | SUGAR | COFFEE | GOLD | TOTAL | % ACERTOS | PRINCIPAIS TERMOS ENCONTRADOS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 120 | **EARN** | **68** | | | 2 | | | | | | | 70 | 56,7 | Qtr, shr |
| 120 | **ACQ** | **60** | 14 | | 18 | 15 | 22 | 40 | 16 | 36 | 13 | 234 | 11,7 | Reuter, Tell |
| 120 | **CRUDE** | | 10 | 4 | 14 | 2 | 2 | **20** | | | | 52 | 3,3 | Money, Market |
| 120 | **TRADE** | | 8 | | 20 | | | 10 | | **36** | 4 | 78 | 16,7 | Plantation, Export |
| 120 | **GRAIN** | | 2 | | | 6 | | 4 | 4 | **8** | | 24 | 5 | Dauster, Ibc, Wheat |
| 120 | **SHIP** | | | | 4 | 2 | **28** | | 2 | 6 | 2 | 44 | 23,3 | Leader, Gob |
| 120 | **MONEY** | | | | 4 | 2 | 8 | **66** | 6 | 4 | | 90 | 55 | Money, Market, Shortage |
| 120 | **SUGAR** | | | | | 4 | | 6 | **14** | | 2 | 26 | 11,7 | Sugar, White |
| 120 | **COFFEE** | | | 12 | 10 | 6 | 4 | | 12 | **114** | | 168 | 95 | Coffee, Quota |
| 120 | **GOLD** | | | 6 | 2 | | | 2 | | | **120** | 130 | 100 | Gold, Mine |
| **1200** | **TAL** | **128** | **34** | **22** | **74** | **37** | **64** | **148** | **54** | **214** | **141** | **916** | **37,8** | |

Tabela 4.30: EC4 – *PA* – Resultados da Clusterização

Conforme mostrado na tabela 4.30, esse é o pior resultado alcançado em todos os EC com essa ferramenta, na tarefa de Clusterização, com apenas 37,8% de acerto.

Os grupos ‘COFFEE’ e ‘GOLD’ obtiveram expressivos resultados em termos de precisão, enquanto que os grupos ‘EARN’ e ‘MONEY’ alcançaram resultados apenas medianos. Os demais obtiveram resultados insatisfatórios, sendo que os grupo ‘CRUDE’ e ‘GRAIN’ tiveram resultados muito ruins.

Vale ressaltar que o método tratou apenas 916 dos 1200 documentos envolvidos, o que corresponde a 76,3% do total.

### 4.3.5. ESTUDO DE CASO 5 (EC5)

Esse EC5 foi elaborado para possibilitar a avaliação das tarefas de Visualização dispostas no *ORACLE TEXT*. Essas tarefas dizem respeito à extração de informações de um único documento, e não de um conjunto, como nas tarefas de Classificação e Clusterização.

Existem cinco possíveis tarefas ligadas a Visualização dispostas no programa. Dessas, três dizem respeito a diferentes apresentações dos resultados com relação ao destaque de termos ou temas. São elas:

- Uma versão do documento com marcações nos termos da consulta;
- Uma versão do documento com termos destacados;
- Uma apresentação do elemento consultado dentro do seu contexto.

As outras 2 são as seguintes:

- Obtenção do Assunto Principal do documento;
- Obtenção da Lista de Temas do documento;

<TITLE>BRAZIL HAS NO SET COFFEE EXPORT TARGETS – IBC </TITLE>
<BODY>Brazil has no set target for its coffee exports following this week´s breakdown of International Coffee Organization talks on export quotas, President of the Brazilian Coffee Institute, IBC, Jorio Dauster said.

He told a press conference Brazil now had to reconsider its export plans and that the 15.5 mln bag export figure which it had proposed for itself earlier should no longer be taken as the country´s export target to ICO-member countries.

The 15.5 mln bag offer had been made on the assumption an agreement would bring stability to world markets, he added. </BODY>

Figura 4.4 : Documento da classe ‘COFFEE’ utilizado nas tarefas de Visualização

Foram selecionados documentos das dez classes integrantes dos EC para a realização de experimentos nas tarefas de Visualização. Para simplificar a apresentação

neste trabalho dos resultados obtidos, são exibidos os resultados de um único documento, considerado piloto, da classe ‘COFFEE’, que é apresentado pela figura 4.4. A seguir estão dispostas as tarefas de Visualização realizadas.

#### 4.3.5.1. VERSÃO DO DOCUMENTO COM MARCAÇÕES NOS TERMOS DA CONSULTA

Na elaboração das *procedures* que executam essa tarefa, é necessário especificar o documento a ser consultado e os termos a serem marcados, além do padrão de marcação desejado.

Essa tarefa foi configurada para exibir trechos do documento piloto com os termos ‘COFFEE’ e ‘EXPORT’ marcados, como resultado. Foi especificado que a marcação dos termos seria no padrão HTML (<B>...</B>). A figura 4.5 apresenta o resultado obtido.

<TITLE>BRAZIL HAS NO SET <B>COFFEE</B> <B>EXPORT</B> TARGETS - IBC</TITLE>
<BODY>Brazil has no set target for its <B>coffee</B> exports following this week´s breakdown of International <B>Coffee</B> Organization talks on <B>export</B> quotas, President of the Brazilian <B>Coffee</B> Institute, IBC, Jorio Dauster said.

He told a press conference Brazil now had to reconsider its <B>export</B> plans and that the 15.5 mln bag <B>export</B> figure which it had proposed for itself earlier should no longer be taken as the country´s <B>export</B> target to ICO-member countries.

The 15.5 mln bag offer had been made on the assumption an agreement would bring stability to world markets, he added. </BODY>

Figura 4.5 : Versão do documento com marcações nos termos da consulta

#### 4.3.5.2. VERSÃO DO DOCUMENTO COM OS TERMOS DESTACADOS

Para a execução dessa tarefa, é necessário especificar o documento a ser utilizado, o padrão de marcação dos elementos e os termos ou temas a serem destacados.

Foi definido que o Tema 'COFFEE' do documento piloto deveria ser destacado, desta vez o Tema e não o termo, e o tipo de marcação especificado foi a apresentação dos elementos com a fonte em negrito e no padrão HTML. A figura 4.6 apresenta o resultado obtido.

<TITLE>BRAZIL HAS NO SET <B>**COFFEE**</B> EXPORT TARGETS - IBC</TITLE>
<BODY>Brazil has no set target for its <B>**coffee**</B> exports following this week´s breakdown of International <B>**Coffee**</B> Organization talks on export quotas, President of the Brazilian <B>**Coffee**</B> Institute, IBC, Jorio Dauster said.

He told a press conference Brazil now had to reconsider its export plans and that the 15.5 mln bag export figure which it had proposed for itself earlier should no longer be taken as the country´s export target to ICO-member countries.

The 15.5 mln bag offer had been made on the assumption an agreement would bring stability to world markets, he added. </BODY>

Figura 4.6 : Versão do documento com os termos destacados

#### 4.3.5.3. APRESENTAÇÃO DO ELEMENTO CONSULTADO DENTRO DO SEU CONTEXTO

Essa tarefa foi configurada para recuperar e exibir o trecho do documento piloto que apresentasse os termos 'EXPORT' e 'PLANS' dentro dos seus contextos. Foi definido também que a marcação dos termos seria no padrão HTML. A figura 4.7 apresenta o resultado obtido.

Brazil now had to reconsider its <B>export</B> <B>plans</B> and that the 15.5 mln bag export figure

Figura 4.7 : Apresentação do elemento consultado dentro do seu contexto

### 4.3.5.4. OBTENÇÃO DO ASSUNTO PRINCIPAL DO DOCUMENTO

Essa tarefa foi configurada para exibir o assunto principal do documento piloto como resultado. Foi especificado que esse assunto deveria ser composto por dois termos e que o resultado deveria ser exibido na tela. O resultado obtido é apresentado pela figura 4.8.

'COFFEE EXPORT'

Figura 4.8 : Apresentação do Assunto Principal do documento

### 4.3.5.5. OBTENÇÃO DA LISTA DE TEMAS DO DOCUMENTO

Essa tarefa foi configurada para extrair do documento piloto uma lista de Temas hierárquica, ou seja, mostrando os relacionamentos e as dependências entre os termos. Foi especificado que o número de Temas retornados deveria ser no máximo 5. Também foi definido que o resultado deveria ser exibido na tela. A figura 4.9 apresenta o resultado obtido.

EXPORT
→ COFFEE
→ BRAZIL

Figura 4.9 : Apresentação dos Temas do documento

O resultado contém apenas três termos, e os dois últimos são dependentes do primeiro.

VISUALIZAÇÃO: ANÁLISE DOS RESULTADOS

É possível observar que o *OT* produziu resultados satisfatórios nas tarefas de Visualização.

As duas tarefas iniciais são bastante simples, e os resultados obtidos estão dentro das expectativas, a partir da definição dos parâmetros das tarefas. Na tarefa de exibição dos elementos dentro dos seus contextos, foi considerado que a sentença recuperada é adequada para contextualizar os termos especificados. Na tarefa de extração do Assunto Principal o resultado obtido se mostrou coerente com o conteúdo do documento e foi considerado satisfatório, de acordo com o especificado na configuração da tarefa. Foi considerado que a ferramenta também obteve êxito na extração da lista de temas do documento e a hierarquia dos termos apresentada no resultado é adequada e coerente com o conteúdo do mesmo.

Cabe assinalar que os experimentos realizados com os documentos das demais classes também obtiveram resultados satisfatórios nas tarefas de Visualização dispostas no *OT*.

# Capítulo V

# CONCLUSÕES

Neste capítulo são discutidos os resultados obtidos nos Estudos de Casos, os critérios de avaliação considerados, a avaliação desses resultados, as dificuldades e limitações encontradas, as estratégias adotadas, os principais detalhes da execução das tarefas, e são apresentadas também as sugestões para futuros trabalhos nessa área.

Na avaliação do desempenho do *ORACLE TEXT* nas tarefas propostas, foram levados em consideração os seguintes aspectos:

- A precisão ou acurácia dos resultados;
- A dificuldade na elaboração, configuração e execução das tarefas;
- As limitações encontradas;
- A apresentação dos resultados obtidos;
- A escalabilidade do sistema;
- O tempo de resposta.

Com relação as tarefas de Classificação e Clusterização, a tabela 5.1 mostra o resumo dos resultados obtidos tanto no *ORACLE TEXT* quanto no *Poly Analyst*, onde é apresentado o percentual médio de acertos nas tarefas indicadas, além da configuração geral dos Estudos de Casos.

| | ESTUDOS DE CASO<br><br>CONFIGURAÇÕES | EC1 | EC2 | EC3 | EC4 |
|---|---|---|---|---|---|
| CATEGORIAS | 1- EARN | 30 | 120 | 30 | 120 |
| | 2- ACQ | 30 | 120 | 30 | 120 |
| | 3- CRUDE | 30 | 120 | 30 | 120 |
| | 4- TRADE | 30 | 120 | 30 | 120 |
| | 5- GRAIN | 30 | 120 | 30 | 120 |
| | 6- SHIP | 30 | 120 | 30 | 120 |
| | 7- MONEY | | | 30 | 120 |
| | 8- SUGAR | | | 30 | 120 |
| | 9- COFFEE | | | 30 | 120 |
| | 10- GOLD | | | 30 | 120 |
| | **TOTAL DE DOCUMENTOS** | **180** | **720** | **300** | **1200** |
| | | | | | |
| ORACLE TEXT | **CLASSIFICAÇÃO BASEADA EM REGRAS** | **100%** | **97,5%** | **100%** | **98,5%** |
| | **CLASSIFICAÇÃO SUPERV. - ÁRVORES DE DECISÃO** | **98,3%** | **97,5%** | **96%** | **95,8%** |
| | **CLASSIFICAÇÃO SUPERV. - ALGORITMO *SVM*** | **100%** | **99,2%** | **98%** | **97,5%** |
| | **CLUSTERIZAÇÃO** | **57,2%** | **59,1%** | **62,3%** | **64,2%** |
| *POLY ANALYST* | **CLASSIFICAÇÃO LINEAR - BAYESIANO SIMPLES** | **98,3%** | **97,5%** | **97%** | **96,5%** |
| | **CLASSIFICAÇÃO LINEAR - ALGORITMO *SVM*** | **100%** | **99,6%** | **98%** | **97,8%** |
| | **CLUSTERIZAÇÃO** | **53,4%** | **50,2%** | **39%** | **37,8%** |

Tabela 5.31: Resumo das tarefas de Classificação e Clusterização, com as % de acerto

É possível concluir que nas tarefas de Classificação as abordagens dispostas pelo *OT* são adequadas para o tratamento de conjuntos de dados textuais, com características e tamanhos variados. Os índices de acerto foram considerados excelentes em todas as abordagens, todos acima dos 95%. Certamente que na Classificação baseada em Regras isso se deve também a qualidade das regras geradas manualmente e carregadas no sistema.

Na Clusterização, os resultados alcançados pelo *OT* foram apenas medianos, o que talvez possa ser creditado às características gerais dos dados utilizados, uma vez que o *Poly Analyst* alcançou resultados inferiores, mesmo considerando que cada ferramenta tem um tratamento específico para essa tarefa. Uma limitação encontrada aqui é o fato do *OT* dispor de apenas uma abordagem para essa tarefa, baseada no algoritmo *K-Means*, o que confere ao usuário a tarefa de definição da quantidade de grupos a ser buscada. Outra limitação percebida é o fato do sistema não dispor de

nenhuma abordagem investigativa da possível quantidade de grupos existente nos dados.

Nas tarefas de Visualização os resultados obtidos foram considerados suficientemente satisfatórios. Uma limitação encontrada foi o sistema não dispor de nenhuma abordagem de Sumarização de textos, além do fato dos resultados serem apresentados em formatos pouco elaborados.

Com relação as dificuldades na elaboração das tarefas, cabe destacar que na Classificação Supervisionada, tanto na abordagem baseada em Árvores de Decisão quanto na baseada no algoritmo *SVM*, elas são realizadas para cada documento individualmente, o que torna a sua execução bastante trabalhosa. Essa seria a principal limitação da ferramenta com relação a Classificação.

Os resultados obtidos na Classificação baseada em Regras são apresentados em uma tabela geral, contendo a classificação de todos os documentos, o que facilita o trabalho de visualização e análise. Nas duas abordagens de Classificação Supervisionada o resultado é individual por documento, o que confere mais trabalho na sua análise.

Os resultados da Clusterização são apresentados em duas tabelas, uma contendo os *Clusters* gerados e a outra contendo a associação dos documentos a eles, o que é considerado bastante adequado para a sua análise.

Os resultados obtidos pelas tarefas de Visualização são de fácil entendimento e considerados adequados, embora de apresentação muito simples.

De uma maneira geral, a elaboração e a configuração dos procedimentos para a execução das tarefas é de complexidade média, considerando que o usuário tenha um conhecimento mínimo no Banco de Dados Oracle.

Em todas as tarefas executadas a ferramenta mostrou ter capacidade de tratar grandes massas de dados, obtendo os resultados com muito pouco tempo de processamento. O principal parâmetro para essa análise diz respeito a Clusterização realizada pelo *OT* no EC4: por um lado, essa é a tarefa que mais requer processamento computacional; por outro, esse é o maior EC elaborado, em que foram utilizados 1200 documentos na Clusterização. Na execução dessa tarefa, o sistema gastou menos de 4 segundos no processamento, o que foi considerado bastante satisfatório, demonstrando a escalabilidade da ferramenta no processamento de grandes conjuntos de dados textuais.

## 5.1 SUGESTÕES PARA TRABALHOS FUTUROS

Nesta seção são apresentadas as sugestões para futuros trabalhos nessa área de pesquisa, visando complementar e expandir os desenvolvidos aqui. Algumas dessas sugestões:

- Novos experimentos com novas bases de dados, contendo grandes volumes de documentos;
- Execução da tarefa de Clusterização em ferramentas que implementem a mesma abordagem que o *OT*, baseada no algoritmo *K-MEANS*, para as mesmas bases de dados utilizadas nos Estudos de Casos, para uma melhor avaliação dos resultados obtidos com o Oracle nessa tarefa;
- Execução de tarefas de Visualização em ferramentas que produzam resultados equivalentes aos produzidas pelo *OT*, para uma avaliação mais aprofundada das tarefas dispostas no *OT*;
- Desenvolvimento de um *Thesaurus* específico para os EC elaborados e nova execução das tarefas de Clusterização, que foram as tarefas que obtiveram os resultados menos expressivos em termos de acurácia, para poder avaliar a contribuição que um dicionário desse tipo agrega ao sistema na abordagem dessa base de dados.

# Referências Bibliográficas


[1] EBECKEN, N. F. F., LOPES, M. C. S., COSTA, M. C. A., "Mineração de textos". In: *Sistemas inteligentes: fundamentos e aplicação*. Barueri Manole. cap. 13, p. 337-370, 2003.

[2] DANTAS, M. A. R., *Implementação de Metodologia de Categorização de Textos Científicos*, Dissertação de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2007.

[3] CAVALCANTI, M., GOMES, E., PEREIRA, E., *Gestão de Empresas na Sociedade do Conhecimento*. 2 ed. Rio de Janeiro, RJ, Brasil, Campus, 2001.

[4] DRUKER, P., *Sociedade Pós-capitalista*, 1 ed. São Paulo, SP, Brasil, Pioneira, 1997.

[5] ALONSO, O., FORD, R., "Text Mining with *ORACLE TEXT*", January 2005. Disponível em <http://www.oracle.com/products/text/>, Acesso em 05 de Setembro de 2007.

[6] GOLDSCHMIDT, R., PASSOS, E., *Data Mining – Um guia prático – Conceitos, técnicas, ferramentas, orientações e aplicações*. 1 ed. Rio de Janeiro, RJ, Brasil, Elsevier, 2005.

[7] FAYYAD, U.M., PIATETSKY-SHAPIRO, G., SMYTH, P., "From Data Mining to knowledge Discovery: an Overview", *Knowledge Discovery and Data Mining International Conference*, Menlo Park, pp. 5-12, 1996.

[8] ZANASI, A., "Web and Text Mining for Open Sources Analysis and Competitive Intelligence". In: *IBM Government Solutions*, Bologna KDD Center, Italy, 2001.

[9] TAN, A., 1999, "Text Mining: the state of the art and the challenges", Disponível em: <http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf>. Acesso em 20 de Outubro de 2007.

[10] LOPES, M. C. S., *Mineração de dados textuais utilizando técnicas de clustering, para o idioma português*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004.

[11] FURTADO, M. I. V., *Inteligência Competitiva para o Ensino Superior Privado: uma Abordagem através da Mineração de Textos*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004.

[12] GONÇALVES, L. S. M., REZENDE, S. O., "Categorização em Text Mining". Disponível em <http://www.di.ubi.pt/~api/text_categorization.pdf>. Acesso em: 02 set. 2007.

[13] NETO, F. L., DINIZ, C. A. R., *Data mining: uma introdução*. 2 ed. São Paulo, SP, Brasil, Associação Brasileira de Estatística, 2000.

[14] KONCHADY, M., *Text Mining Application Programming*, 1 ed. Boston, Massachusetts, EUA, Charles River Media, 2006.

[15] WEISS, S. M., INDURKHYA, N., ZHANG, T., et. al. *Text Mining – Predictive Methods for Analyzing Unstrutured Information*, 1 ed. New York, NY, EUA, Springer Science Business Media, 2005.

[16] HAN, J., KAMBER, M., *Data Mining – Concepts and Techiniques*, 1 ed. San Diego, CA, USA, Morgan Kaufmann Publishers, 2001.

[17] HAHN, U., MANI, I., "The challenges of Automatic Summarization". *IEEE Computer*, Vol. 33, No. 11, November, 2000.

[18] SUN, A., LIM, E., "Hierarchical Text Classification and Evaluation", In: *IEEE International Conference on Data Mining*, C.N.R., California, USA, p.521-528, 2005.

[19] SEBASTIANI, F., "Machine learning in automated text categorisation: a survey", *Technical Report IEI-B4-31-1999*, Istituto di Elaborazione dell'Informazione, C.N.R., Pisa, Italia, 1999.

[20] MARTINS, C. J. M., *Aplicação de Ferramentas Computacionais para Prospecção Tecnológica por Mineração de Dados Não Estruturados sobre*

*Patentes Industriais em Idiomas Inglês*, Dissertação de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2008.

[21] CALINSKI, R.B., HARABASZ, J., *A dendrite method for cluster analysis*. In: Communication in statistics, vol. 3, pp. 1-27, 1974.

[22] PAKHIRA, M. K., BANDYOPADHYAY, S., MAULIK, U., *Validity index for crisp and fuzzy cluster*, Pattern Recognition, vol. 37, no. 3 (Mar), pp. 487-501, 2004.

[23] XIE, X., BENI, G., *A validity measure for fuzzy clustering*. In: IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 13, n. 8, pp. 841-847, 1991.

[24] ORACLE, *ORACLE TEXT – Application Developer´s Guide 10g, Release 2* (10.2), June 2005. Disponível em <http://www.oracle.com/technology/products/text/ index.html>

[25] ORACLE, *ORACLE TEXT – Reference 10g, Release 2 (10.2)*, June 2005. Disponível em <http://www.oracle.com/technology/products/text/index.html>

[26] ORACLE, *Oracle Application Server Portal – Configuration Guide 10g, Release 2 (10.2)*, November 2005. Disponível em <http://www.oracle.com/technology/ products/text/index.html>

[27] ORACLE, *Oracle Data Mining – Administrator´s Guide 10g, Release 2 (10.2)*, November 2005. Disponível em <http://www.oracle.com/technology/products/ text/index.html>

[28] MATSUNAGA, L. A., *Uma Metodologia de Categorização Automática de Textos para a Distribuição dos Projetos de Lei às Comissões Permanentes da Câmara Legislativa do Distrito Federal*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2007.

[29] TOAD, site na internet: Toad® - *Tools for Database Development & Administration*, Disponível em <http://www.quest.com/toad-for-oracle>, Acesso em 10 de Outubro de 2007.

[30] *POLY ANALYST, Poly Analyst Help, Poly Analyst 6* – Megaputer Intelligence Inc. – Tutorial do Software, 2007.

[31] OREN, Z., 1999, *Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results*, Ph.D. dissertation, University of California at Berkeley, Berkeley, California, USA.

# Livros Grátis

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

Milhares de Livros para Download:

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