COPPE/UFRJ

MINERAÇÃO DE TEXTOS COM GEOREFERENCIAMENTO

Ronaldo Braga Lopes

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do Título de Doutor em Engenharia Civil

Orientador: Alexandre Gonçalves Evsukoff

Rio de Janeiro
Dezembro de 2009

MINERAÇÃO DE TEXTOS COM GEOREFERENCIAMENTO

Ronaldo Braga Lopes

TESE SUBMETIDA AO CORPO DOCENTE DO INSTTUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

____
Prof. Alexandre Gonçalves Evsukoff, Dr.

____
Prof. Nelson Francisco Favilla Ebecken, D.Sc.

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

____
Prof. Elton Fernandes, Ph.D.

____
Prof. Paulo Cezar Pinto Carvalho, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

DEZEMBRO DE 2009

Lopes, Ronaldo Braga

Mineração de Textos com Georeferenciamento / Ronaldo Braga Lopes. – Rio de Janeiro: UFRJ/COPPE, 2009.

XIII, 136 p.: il.; 29,7 cm.

Orientador: Alexandre Gonçalves Evsukoff

Tese (doutorado) – UFRJ / COPPE / Programa de Engenharia Civil, 2009.

Referências Bibliográficas: p. 115-123.

1. Mineração de Dados. 2. Mineração de Textos e Georeferenciamento. I. Evsukoff, Alexandre Gonçalves. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

Dedicatória:

À Deus, pela oportunidade de estudar, de aprender.

À minha filha, Maria Gabriela, Luz de minha Vida.

À minha esposa, Alexandra, por seu Amor e incentivo nos momentos de desânimo.

Agradecimentos:

Fico muito feliz em poder dizer que, com muito esforço, consegui conciliar a atividade acadêmica com as responsabilidades profissionais ao longo destes anos de estudo. Tudo isto foi possível, pois tive muitos colegas e amigos que me ajudaram a chegar até aqui. Em especial agradeço:

- ao meu Orientador, pelo incentivo, confiança e dedicação

- à minha esposa por não me deixar desistir, nunca

- à minha família pela compreensão nas horas em que estive ausente

- à minha mãe, pela minha formação escolar e formação de caráter

- ao SAS pela parceria e pela disponibilização de uma Licença Acadêmica

- à Embratel, pelo incentivo à execução deste trabalho

- à COPPE, pela excelência de seus Professores

A todos, Muito Obrigado !!

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

MINERAÇÃO DE TEXTOS COM GEOREFERENCIAMENTO

Ronaldo Braga Lopes

Dezembro / 2009

Orientador: Alexandre Gonçalves Evsukoff

Programa: Engenharia Civil

Este trabalho evidencia um processo de identificação de percepções de consumidores através da utilização de recursos de Mineração de Textos, Mineração de Dados e de SIG's. A tese evidencia na pesquisa bibliográfica os aspectos mais importantes destes assuntos para então, como principal contribuição, detalhar a integração de algoritmos e ferramentas que permitam, a partir de uma base de textos, identificar clusters, suas respectivas palavras-chave e as respectivas correlações entre termos representativos. O processo ainda permite a identificação de características sócio-econômico-demográficas a partir da construção de mapas temáticos em um SIG. Finalizando o trabalho, um Estudo de Caso, evidencia a aplicabilidade da proposta apresentada.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

TEXT MINING WITH GEOREFERENCING

Ronaldo Braga Lopes

December/2009

Advisor: Alexandre Gonçalves Evsukoff

Department: Civil Engineering

This work presents a process aimed at identifying customer's perceptions through the application of Data Mining, Text Mining and Georeferencing tools. This thesis presents bibliographic research analyzing the most important issues of these subjects, and then, as a main contribution, details the integration of algorithms and tools that allow, from a textual database, obtain the clusters identification, their keywords and the correlations between the most representative terms. Moreover, the process permits the identification of the social-economic-demographical characteristics from the construction of thematic maps in a GIS. Finally the work presents a Case Study giving evidences of the applicability of presented proposal.

ÍNDICE

CAPÍTULO 1: O PROBLEMA DE PESQUISA ... 1

1.1 INTRODUÇÃO ... 1

1.2 DEFINIÇÃO DO PROBLEMA ... 1

1.3 JUSTIFICATIVA ... 2

1.4 CONTRIBUIÇÕES DA PESQUISA ... 2

1.4.1 Objetivo Geral ... 3

1.4.2 Objetivos Específicos ... 3

1.4.3 Originalidade ... 4

1.5 ESTRUTURAÇÃO DO TRABALHO ... 4

CAPÍTULO 2: MINERAÇÃO DE TEXTOS ... 5

2.1 O PRÉ-PROCESSAMENTO DOS DADOS ... 6

2.1.1 Stop Words, Thesaurus, Stemmers ... 6

Eliminação de *Stop Words* ... 7

Thesaurus ... 7

Stemming ... 7

2.1.2 Representação de Documentos ... 9

2.1.3 Redução de Dimensionalidade ... 11

2.2 O PROCESSAMENTO ... 13

2.2.1 Clusterização de Documentos ... 13

Algoritmo K-Means ... 14

Clusterização Hierárquica ... 15

O Método “Ward” de Clusterização Hierárquica ... 16

O Algoritmo de Expectation-Maximixation (EM) ... 17

2.2.2 Classificação de Documentos ... 21

Conceitos Básicos ... 21

O Classificador Bayesiano ... 22

Support Vector Machine (SVM) ... 22

2.3 OPINION MINING ... 24

2.3.1 Classificação do Tipo de Sentimento ... 24

2.3.2 POS: Part of Speech Tagging (POS) ... 25

2.3.3 Classificação de Opinião utilizando Métodos Convencionais ... 29

2.4 RESUMO DO CAPÍTULO ... 30

CAPÍTULO 3: SISTEMAS DE INFORMAÇÃO GEOGRÁFICA ... 31

3.1 CONCEITOS BÁSICOS DE ESPAÇO GEOGRÁFICO E DE INFORMAÇÃO ESPACIAL ... 31

3.1.1 Relações Espaciais nos Fenômenos Geográficos ... 32

3.2 MODELOS DE DADOS PARA GEOPROCESSAMENTO ... 33

3.3 MODELAGEM ESPACIAL: ... 34

3.3.1 Conceitos Básicos de Modelagem Geoestatística ... 34

Teoria das variáveis regionalizadas 35
Relação Espacial 36
Semivariograma 36
Krigagem 39
3.4 SOFTWARES PARA ANÁLISE ESPACIAL 40
3.7 RESUMO DO CAPÍTULO 44
CAPÍTULO 4: METODOLOGIA DE PESQUISA 45
4.1 VISÃO GERAL DA METODOLOGIA: 45
4.2. A ETAPA DE MINERAÇÃO DE TEXTOS 46
4.1.1 Etapa A: Extração e Limpeza Inicial da Base de Dados 47
1. Extração dos Dados 48
2. Conversão para Arquivos Manipuláveis 48
3. Limpeza Bruta 48
4. Filtro por Características 48
5. Redução (1 arquivo – 1 linha) 49
6. Criação de Biblioteca e "*Data Source*" 49
4.1.2 Etapa B: Pré-Processamento dos Dados 50
7. Modelagem Vetorial 50
8. Correção de "*Mispelling*" 51
9. Stemming 53
10. Aplicação de *Part Of Speech Tagging* (*POS Tagging)* 54
11. Redução de "*Stop Words*" 54
13. LSI 55
14. Clusterização (*Expectation – Maximization*) 57
15. Avaliação dos Resultados Estatísticos 57
16. Termos Representativos 57
17. Análise de Correlação 58
18. Novas Tentativas de *Clustering* 59
19. Análise Descritiva dos Clusters Obtidos 59
4.3 A ETAPA DE GEOREFERENCIAMENTO 59
4.3.1 Detalhamento do Processo de Georeferenciamento 60
20. Preparação da Camada de Referência 61
21. Georeferenciamento dos Documentos Textuais 63
22. Georeferenciamento de Informações Complementares 65
23. Consolidação das Camadas 66
24.Mapas Temáticos dos Clusters 66
25. Google Maps / Google Earth 67
4.4 RESUMO DO CAPÍTULO 67
CAPÍTULO 5: O ESTUDO DE CASO 68
5.1 O BANCO DE DADOS TEXTUAL 68
5.2 O PRÉ –PROCESSAMENTO DA BASE 69
A Geração de Stop Words 71

Thesaurus 73
O Processo de *Stemming* 74
A Atribuição de Pesos aos Termos 75
A Redução de Dimensionalidade 77
5.3 O PROCESSAMENTO DOS TEXTOS 78
Conceitos Extraídos dos Clusters Obtidos 81
5.3.1 Descrição / Resumo dos Clusters Obtidos 82
Cluster 1: Reparos 82
Cluster 2: Problema de Logística de Entrega (Aparelho e Fatura de Clientes) 84
Cluster 3: Solicitação de Segunda Via / Boleto com Código de Barras para Pagamento 85
Cluster 4: Solicitações de Migração de Saldo de um Ciclo (de Pagamento para Outro) 86
5.4 O PROCESSO DE GEOREFERENCIAMENTO 88
5.4.1 A Representação da Área de Cobertura 89
O Georeferenciamento dos Clusters 98
Considerações sobre o Cluster 1 (Reparos) 105
Considerações sobre o Cluster 2 (Problema de Logística de Entrega de Aparelho e Fatura de Clientes).. 108
Considerações sobre o Cluster 3 (Solicitação de Segunda Via) 110
Considerações sobre o Cluster 4 (Solicitações de Migração de Saldo) 111
Considerações sobre o Cluster 5 (Questionamentos sobre o Valor da Conta) 112
5.5 RESUMO DO CAPÍTULO 112

6. CONCLUSÕES E SUGESTÕES PARA FUTURAS PESQUISAS 114

REFERÊNCIAS BIBLIOGRÁFICAS 116

APÊNDICE A: 125

APÊNDICE B: 126

APÊNDICE C: 127

APÊNDICE D: 127

APÊNDICE E: 127

APÊNDICE F: 127

APÊNDICE G: 132

ÍNDICE DE FIGURAS

FIGURA 1: REPRESENTAÇÃO DE DOCUMENTOS NO MODELO VETORIAL 9
FIGURA 2: EXEMPLO GRÁFICO DE CLUSTERIZAÇÃO UTILIZANDO O K-MEANS 20
FIGURA 3: WORKFLOW DO CLUSTERING POR EXPECTATION-MAXIMIZATION 20
FIGURA 4: REPRESENTAÇÃO DO HIPERPLANO 23
FIGURA 5: OS QUATRO UNIVERSOS SEGUNDO MONTEIRO E CÂMARA (2007) 33
FIGURA 6: VARIÁVEL ALEATÓRIA REGIONALIZADA Z (X) 33
FIGURA 7: SEMIVARIOGRAMA AJUSTADO 38
FIGURA 8: MACRO-PROCESSO DO SISTEMA PROPOSTO 45
FIGURA 9: RESUMO DA METODOLOGIA DA ETAPA DE MINERAÇÃO DE TEXTOS 46
FIGURA 10: PROCESSO DE EXTRAÇÃO E LIMPEZA BRUTA DOS DADOS 47
FIGURA 11: PRÉ-PROCESSAMENTO DOS DADOS 50
FIGURA 12: WORKFLOW PARA CRIAÇÃO DE LISTA DE SINÔNIMOS 51
FIGURA 13: WORKFLOW DA ETAPA DE PROCESSAMENTO 56
FIGURA 14: EXEMPLO DE CAMADAS DE REPRESENTAÇÃO EM UM SIG 60
FIGURA 15: WORKFLOW DO PROCESSO DE GEOREFERENCIAMENTO 60
FIGURA 16: GEOREFERENCIAMENTO BASEADO NO CEP 64
FIGURA 17:EXEMPLO DE REPRESENTAÇÃO TEMÁTICA 66
FIGURA 18: AMOSTRA DO DATASET ORIGINAL 69
FIGURA 19: LOG DE GERAÇÃO DE ARQUIVO TEXTUAL PARA ETAPA DE PRÉ-PROCESSAMENTO 70
FIGURA 20: AMOSTRA DO BANCO DE DADOS TEXTUAL 71
FIGURA 21: LISTAGEM DE PALAVRAS ADICIONADAS À STOP WORDS LIST 72
FIGURA 22: AMOSTRA DO *CORPUS* 73
FIGURA 23: LISTA DE SINÔNIMOS APLICADA AO DATASET ORIGINAL 74
FIGURA 24: EXEMPLO DE ALGORITMO DE STEMMING APLICADO AO DATASET 75
FIGURA 25: PLOTAGEM DAS FREQUÊNCIAS X QUANTIDADE DE DOCUMENTOS 76
FIGURA 26: ATRIBUIÇÃO DE PESOS COM O MÉTODO IDF 76
FIGURA 27: PESOS ATRIBUÍDOS A CADA TERMO 77
FIGURA 28: AMOSTRA DA MATRIZ SVD GERADA 78
FIGURA 29: DISTRIBUIÇÃO DA QTD. DE DOCUMENTOS P/ CLUSTER (NA CLUSTERIZAÇÃO INICIAL) 79
FIGURA 30: LOCALIZAÇÃO RELATIVA DOS CENTRÓIDES (NO PROCESSO INICIAL DE CLUSTERIZAÇÃO) 80
FIGURA 31: LOCALIZAÇÃO DOS CENTRÓIDES DOS CLUSTERS FINAIS 81
FIGURA 32: CORRELAÇÃO ENTRE OS TERMOS MAIS EXPLICATIVOS DO CLUSTER 1 83
FIGURA 33: CORRELAÇÃO ENTRE OS TERMOS MAIS EXPLICATIVOS DO CLUSTER 2 84
FIGURA 34: CORRELAÇÃO ENTRE OS TERMOS MAIS EXPLICATIVOS DO CLUSTER 3 85
FIGURA 35: CORRELAÇÃO ENTRE OS TERMOS MAIS EXPLICATIVOS DO CLUSTER 4 86
FIGURA 37: CORRELAÇÃO ENTRE OS TERMOS MAIS EXPLICATIVOS DO CLUSTER 5 87
FIGURA 38: PASSO 1DE 2 PARA ATUALIZAÇÃO DE TABELA 90
FIGURA 39: PASSO 2 DE 2 PARA ATUALIZAÇÃO DE TABELA 90

FIGURA 40: CAMADA: BAIRROS DA REGIÃO METROPOLITANA DO RJ 95
FIGURA 41: ÁREA DE COBERTURA 95
FIGURA 42: MAPA DE RUAS 95
FIGURA. 43: REPRESENTAÇÃO DA CAMADA DE RUAS COM *"ZOOM"* 95
FIGURA 44: REPRESENTAÇÃO DA SUPERPOSIÇÃO DE CAMADAS 95
FIGURA 45:REPRESENTAÇÃO DA CAMADA DE CLASSES SOCIAIS 95
FIGURA 46: DETALHAMENTO ("ZOOM") DE UMA ÁREA DA ZONA SUL DO RIO DE JANEIRO 98
FIGURA 47: CLUSTER 1: REPAROS 100
FIGURA 48: CLUSTER 2: PROBLEMAS DE LOGÍSTICA DE ENTREGA 101
FIGURA 49: CLUSTER 3:SOLICITAÇÕES DE SEGUNDA VIA / CÓDIGO DE BARRAS 102
FIGURA 50: CLUSTER 4:SOLICITAÇÕES DE MIGRAÇÃO DE SALDO 103
FIGURA 51: CLUSTERR 5: QUESTIONAMENTOS SOBRE SALDO DE CONTAS 104
FIGURA 52: DETALHAMENTO DE ÁREA CRÍTICA DO CLUSTER 1(REPAROS) 106
FIGURA 53: DETALHAMENTO DE RUAS CRÍTICAS DO CLUSTER 1 (REPAROS) 107
FIGURA 54: DETALHAMENTO DE UMA DAS ÁREAS CRÍTICAS DO CLUSTER 2 108
FIGURA 55: FOTO DE SATÉLITE DE REGIÃO CRÍTICA DO CLUSTER 2 109
FIGURA 56: VIZUALIZAÇÃO DE PARTE DA ÁREA DE COBERTURA C/ MENOR CONCENTR. SEG. VIA 111
FIGURA 57: DETALHAMENTO DE ÁREA DA CIDADE COM MAIOR CONCENTRAÇÃO DE CLASSE AB 111

ÍNDICE DE TABELAS

TABELA 1: REPRESENTAÇÃO DOS DOCUMENTOS DO CORPUS 9
TABELA 2: TAGS PARA EXTRAÇÃO DE BINÔMIO DE PALAVRAS 26
TABELA 3: EXEMPLO DE APLICAÇÃO DE ORIENTAÇÃO SEMÂNTICA 29
TABELA 4: EXEMPLO DE APLICAÇÃO DE ORIENTAÇÃO SEMÂNTICA 29
TABELA 5: CUSTOS DAS OPERAÇÕES DO ALGORITMO SPEDIS 52
TABELA 6: ESTATÍSTICAS DOS PRIMEIROS CLUSTERS ENCONTRADOS 80
TABELA 7: PALAVRAS-CHAVE DOS CLUSTERS FINAIS 82

CAPÍTULO 1: O PROBLEMA DE PESQUISA

1.1 INTRODUÇÃO

No mundo atual, é crescente a velocidade do fluxo da informação e cada vez maior a disseminação desta para os consumidores. Neste contexto, é cada vez mais difícil para as empresas e instituições entenderem as opiniões de seus públicos-alvo, de maneira abrangente e na velocidade com que estas mudanças acontecem.

Segundo Amaral (1998, p17):

> *"A informação é um fator imprescindível para impulsionar o desenvolvimento da sociedade, constituindo-se em um insumo de fundamental importância de geração de conhecimento que, por sua vez, possibilitará, de modo eficiente, a satisfação das diversas demandas da população."*

A fim de conseguir atender a esta demanda e atuar sobre estas mudanças é desejável que exista um processo de análise das informações destes consumidores obtidas diretamente na Web ou de outros meios eletrônicos disponíveis nos dias de hoje. Contudo um processo de análise desta natureza requer o conhecimento de especialistas e consome recursos e tempo significativos. Dado que estas informações podem ser extremamente relevantes e que existem incontáveis formas destes consumidores expressarem livremente suas opiniões por este meio, sejam elas positivas ou negativas, o estabelecimento de um processo automatizado de busca e análise de percepções de consumidores, sem dúvida nenhuma pode agregar valor às organizações como uma importante ferramenta de Inteligência Competitiva ou de melhoria de processos / produtos / produtos.

1.2 DEFINIÇÃO DO PROBLEMA

Conforme dito anteriormente, no mundo globalizado, informação é vantagem competitiva. Empresas e organizações atualmente vivenciam um estágio de competição sem precedentes e um dos diferenciais competitivos é o entendimento das percepções dos consumidores em tempo hábil. Entender aquilo que os consumidores pensam de uma organização ou de um produto é um diferencial para o correto posicionamento de uma empresa ou de uma marca.

A maioria das empresas dispõe de farto material de informação não estruturada (textos) tanto na Web quanto em diversos Sistemas de Atendimento a Clientes. Contudo, estas organizações poderiam ter maior eficiência ao conseguir qualificar adequadamente as opiniões de seus consumidores. A dificuldade, ou melhor, o problema para execução de um processo desta forma está no fato de que cada depoimento (de um consumidor) precisa ser analisado, comparado com os demais; para ao final do processo, conseguir-se extrair conceitos que possibilitem a tomada de decisões.

1.3 JUSTIFICATIVA

Buscar entender de maneira estruturada aquilo que os clientes estão pensando de uma empresa, um produto, um serviço, é uma estratégia coerente e que permite a fidelização destes (clientes).

Segundo a pesquisa do *Technical Assistance Research Program* (CZINKOTA, 2001, p.278), foi verificado que:

- em média, as empresas não ficam sabendo de 96% dos clientes insatisfeitos;
- para cada queixa recebida, outros 26 clientes têm o mesmo problema;
- uma pessoa, em média, com um problema fala para nove ou 10 pessoas;
- treze por cento dos clientes dizem a mais de 20 pessoas;
- os clientes que têm seus problemas resolvidos satisfatoriamente contam a uma média de cinco pessoas sobre o tratamento recebido

Desta forma, conseguir entender, através de um processo estruturado, aquilo que os clientes estão manifestando espontaneamente através de relatos escritos, é uma forma de permitir que o fluxo de informação (destas opiniões), chegue de maneira eficiente também à empresa. Aliado a isto, melhor ainda, se houver uma indicação de área geográfica para cada um destes relatos, permitindo às empresas saber se determinadas reclamações / solicitações são diferenciadas em função de suas respectivas localizações por bairros, cidades, localidades, etc.

1.4 CONTRIBUIÇÕES DA PESQUISA

A principal contribuição é estruturar um processo para qualificar opiniões de consumidores através da utilização de técnicas e algoritmos de Mineração de Textos e de Geoprocessamento, ajudando desta forma a solucionar o problema inerente à necessidade de entendimento e de qualificação das vontades, dos anseios, dos consumidores.

1.4.1 Objetivo Geral

Desenvolvimento de metodologia para obtenção de informação relevante que permita inferir conceitos a respeito dos relatos de consumidores sobre um determinado produto ou serviço - através de técnicas combinada de Processamento de Dados Textuais e de Georeferenciamento.

1.4.2 Objetivos Específicos

Conseguir a partir da metodologia proposta:

- integrar ferramentas (de software) de Mineração de Dados, Mineração de Textos e de Sistemas de Informação Geográfica em um *workflow* de processamento;

- avaliar algoritmos para "*otimização*" dos processos mencionados acima

- efetuar estudo de caso com a utilização da metodologia proposta

- avaliar o comportamento das percepções de consumidores (sob a forma de dados não estruturados)

- verificar as correlações entre os dados dos consumidores analisados com:

características sócio-econômicas

características referentes às suas respectivas localizações geográficas

1.4.3 Originalidade

Até o momento não foi encontrada, na pesquisa bibliográfica de artigos e revistas científicas especializadas (consultadas), menção a uma metodologia de Mineração de Textos combinada a Georeferenciamento.

1.5 ESTRUTURAÇÃO DO TRABALHO

Nos capítulos seguintes estão discorridos os temas que permitem fundamentar e testar a metodologia de pesquisa proposta. Estão estruturados da seguinte forma:

- Capítulo 2: revisão de literatura referente aos temas direta ou indiretamente ligados a Mineração de Textos
- Capítulo 3: revisão da literatura relacionada à Modelagem Geoestatística, Georeferenciamento e a Sistemas de Informação Geográfica.
- Capítulo 4: metodologia de pesquisa proposta
- Capítulo 5: Estudo de Caso
- Capítulo 6: Conclusões e sugestões para pesquisas posteriores

CAPÍTULO 2: MINERAÇÃO DE TEXTOS

A partir dos anos 80, com o aumento exponencial do volume de informações armazenadas pelas empresas, surgiu a necessidade de extração, de maneira automatizada, de informações relevantes e de padrões de comportamento em grandes massas de dados. Esta nova área da Tecnologia da Informação passou a ser conhecida como KDD – *Knowledge Discovery in Database* (Descoberta de Conhecimento a partir de uma Base de Dados).

Processos de KDD conseguem obter como resultados, informações não triviais contidas nas bases de dados analisadas – de forma diferenciada dos métodos de extração de informação tradicional. Além disso, o KDD é um processo interativo, em que é natural e possível percorrer diversas etapas e eventualmente retornar a alguma etapa anterior em função de descobertas realizadas ao longo do processo de descoberta de informações relevantes.

O objetivo final de um processo de KDD normalmente é a tomada de decisão; baseando-se nas informações relevantes, fruto da descoberta desse processo de Mineração (de Dados).

O processo de KDD pode ser considerado como uma ciência interdisciplinar, pois constantemente recebe contribuições e permite aplicações em diversas áreas. Atualmente é possível encontrar com certa facilidade artigos acadêmicos e aplicações comerciais para as mais diversas áreas, tais como: bioinformática, telecomunicações, seguros, marketing, medicina entre outras.

Mineração de Textos (do inglês "*Text Mining*") é uma das formas de KDD, ou ainda, o processo de extração de informação não trivial de uma coleção de textos, em geral sob a forma de arquivos eletrônicos.

O objetivo central da Mineração de Textos é analisar informação de bases de dados estruturadas ou semi-estruturadas, conseguir processá-las e efetivamente conseguir descobrir valor útil.

Nasukawa e Nagano (2001) dividem o processo de Mineração de Textos em três partes principais:

a. Extração de Conceito: é a obtenção de palavras-chave de um determinado documento. Estas palavras-chave conseguem distinguir um documento em particular de outro que se queira comparar. Por esse motivo, entende-se que a maneira como são feitas estas escolhas (que devem ou não definir o conceito de um documento) seja um importante passo para esta etapa de extração;

b. Aplicação de técnicas de *Data Mining* para extração de regras e modelos: após extrair conceitos apropriados, vários métodos estatísticos podem ser utilizados com as palavras-chave obtidas na primeira etapa. Além disso, podem-se utilizar funções simples para a descoberta de padrões não intuitivos;

c. Análise interativa e gráfica: como o objetivo de oferecer aos usuários diferentes pontos de vista de cada documento e também em função da possível falta de precisão de sistemas de processamento de dados, a possibilidade de interação gráfica torna-se desejável e uma importante parte da análise dos dados obtidos.

2.1 O PRÉ-PROCESSAMENTO DOS DADOS

2.1.1 *Stop Words, Thesaurus, Stemmers*

Para a obtenção dos conceitos (Nasukawa e Nagano, 2001) faz-se necessário a aplicação de uma série de técnicas que permitam eliminar dentro de cada texto as redundâncias e/ou variações morfológicas, ou seja, conseguir transformar o conjunto de documentos em uma base mais limpa, onde o trabalho de representação de documentos, o respectivo processamento dos dados e a consequente interpretação destes, possam ser feitas de maneira mais rápida e eficiente.

A seguir são descritos alguns conceitos e técnicas que são normalmente utilizadas em processos de preparação de dados para a efetiva mineração de dados textuais:

Eliminação de *Stop Words*

No processo de análise de dados, verifica-se que existem certas palavras que ocorrem com muita freqüência e que não agregam conteúdo ao texto analisado.

Existe uma relação ligeiramente linear entre a freqüência das palavras e sua importância para o entendimento do contexto das informações. Palavras com freqüência muito elevada no *dataset* analisado podem ser descartadas, pois não agregam valor ao entendimento do(s) texto(s) analisado. (Manning & Schütze, 1999)

De maneira análoga, palavras com freqüência extremamente baixa podem também ser removidas, pois dada a sua baixa ocorrência, acabam não bem representando os documentos que estejam sendo analisados.

Tanto os casos de frequência excessiva, como os de frequência muito baixa, podem ser considerados como casos de *Stop Words*, que precisam desta forma ser eliminados no processo de análise de dados.

Thesaurus

Uma outra necessidade para a etapa de preparação dos dados textuais é a utilização de um Dicionário de Termos ou *Thesaurus*. O *Thesaurus* é uma lista de termos com seus respectivos sinônimos, hierarquias e respectivos relacionamentos entre si - que ajuda os usuários de um sistema de Recuperação da Informação ou de uma aplicação de Mineração de Textosa encontrar a informação desejada, com menor esforço computacional.

Stemming

Outra questão importante para a etapa de pré-processamento é o tratamento das variações morfológicas das palavras. Estas variações morfológicas normalmente possuem interpretações semânticas similares. Por esse motivo já existem desenvolvidos os chamados Algoritmos de *Stemming* (ou *Stemmers*). Estes algoritmos têm o objetivo de reduzir as palavras aos seus respectivos *stems*, ou seja – reduzir à menor forma morfologicamente equivalente, permitindo dessa forma um melhor custo-benefício de processamento dos textos nesta etapa da Mineração de Textos.

A seguir a descrição de alguns algoritmos de *Stemming,* mais comumente utilizados em aplicações de Mineração de Textos/ Recuperação da Informação:

Stemmer de Porter

Basicamente, o algoritmo de Porter (Porter M., 1980) realiza a remoção de sufixos. A remoção de sufixos é especialmente interessante para aplicações da área de Recuperação da Informação, onde exista uma coleção de documentos referenciados por seus títulos respectivos e/ou seus respectivos resumos (*abstracts*). Ignorando-se o sentido destas palavras, pode-se dizer que um documento é representado por um vetor de palavras ou de termos. O algoritmo pressuopõe que termos com um radical comum possuem significados similares.

Exemplo: **CONECT**AR
CONECTADO
CONECTANDO

Stemmer de Lovins

Este algoritmo também é bastante utilizado (Lovins, J. B., 1968). Este Stemmer remove no máximo um único sufixo por palavra, com apenas uma única passagem por palavra pelo algoritmo. O Stemmer original de Lovins para a língua inglesa possui 250 sufixos diferentes e durante o processo, o sufixo mais longo é retirado da palavra, de forma que o "*stem*" resultante após a remoção tenha pelo menos três caracteres. O stemmer de Lovins tem o problema de querer resolver problemas de Recuperação da Informação e de Lingüística ao mesmo tempo e desta forma não conseguir obter alta performance em ambos. Na verdade, este algoritmo não é complexo o suficiente para extirpar uma série de sufixos que não estejam na lista de regras – uma vez que o mesmo foi criado a partir de um determinado conjunto de exemplos. Possivelmente se este processo tivesse sido realizado com uma amostra muito maior, seria possível obter um stemmer de maior eficácia.

2.1.2 Representação de Documentos

Uma etapa usual no processo de Mineração de Textos é a representação vetorial, onde os documentos são dispostos sob a forma de vetores (vetores de atributos). Para isso, cada dimensão do vetor é correspondente a um único termo do dicionário estipulado, conforme exemplo na Tabela 1. Este tipo de representação é evidenciado na Tabela 1 também conhecido como BOW (*Bag of Words*).

Termo/ Documento	Term 1	Term 2	Term 3	Term 4	Term 5	...	Term n-1	Term n
Doc 1	0	0	1	1	0	...	0	0
Doc 2	0	1	0	0	1	...	0	1
Doc 3	0	1	0	0	0	...	1	1
Doc 4	1	0	1	0	1	...	1	1
Doc 5	1	0	1	0	1	...	0	1
Doc 6	1	1	1	1	0	...	1	0
Doc 7	1	1	0	1	1	...	0	0
...	...	...	...	...	...	...	...	...
Doc n-1	1	0	0	1	0	...	1	1
Doc n	0	1	0	0	1	...	0	0

Tabela 1: Representação dos documentos do Corpus

Esta forma de representação de documentos é conhecida como Modelo de Espaço Vetorial, proposta originalmente por Salton (1975). Esta técnica já é muito utilizada em aplicações de Recuperação da Informação, desde a década de 70. Na Figura 1, uma representação de alguns documentos em um espaço vetorial.

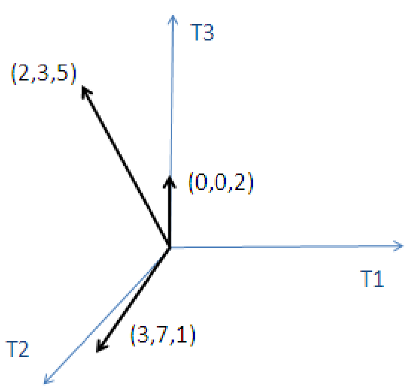


Figura 1: Representação de Documentos no Modelo Vetorial

Seguindo esta linha de raciocínio, termos que possuem alta freqüência têm importância maior em um documento. Por sua vez, se estes termos aparecem em uma grande quantidade de documentos, terão sua importância diminuída. Desta forma, uma importante tarefa para a caracterização da coleção de documentos examinados é a análise da freqüência de palavras na coleção. Este tipo de análise permite verificar a importância de cada palavra em relação aos textos analisados, bem como permite a distinção dos documentos entre si.

De forma a diferenciar os termos em função de sua importância relativa, são atribuídos pesos (aos termos). Os valores dos elementos de um vetor - representação de cada documento através das respectivas frequências de cada termo - são calculados como sendo a combinação das estatísticas TF(t,d) e DF(t). Sendo TF(t,d) o número de vezes que um termo "t" ocorre em um documento "d", enquanto DF(t) é o número de documentos em que tal termo aparece pelo menos uma vez.

Uma maneira usual de se atribuir estes pesos adequadamente é utilizar o cálculo da Freqüência Inversa de um Documento, que é calculada da seguinte forma:

$$IDF = \log (|D| / DF(t))$$

Onde: $|D|$ é o número total de documentos.

Pela simples observação da fórmula pode-se verificar que quando a IDF de uma palavra ou termo é baixa, este termo ocorre em muitos documentos. Conseqüentemente IDF tem seu valor máximo quando o termo aparece em apenas um único documento.

A estatística TF-IDF combina as duas anteriores:

$$TF\text{-}IDF(t) = TF(t,d).\ IDF(t)$$

Desta forma de atribuição de peso decorre que: se um termo Ti ocorre freqüentemente no documento, este é um termo importante para a identificação deste documento. De outra forma, palavras que ocorrem em muitos documentos são pouco importantes para indexação. Em suma, pode-se concluir que a IDF serve como uma função de ajuste, de modulação da freqüência de termos relevantes em um documento.

Para a verificação de similaridade entre documentos primeiramente são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação (conforme Figura 1). Para o cálculo da similaridade propriamente dita é calculado o ângulo entre os vetores. Na prática trabalha-se com o cálculo do cosseno deste ângulo. A título de exemplo, um valor de cosseno igual a zero indica que os vetores são ortogonais e portanto não tem qualquer similaridade.

Outros métodos relevantes para verificação de similaridade são: SWC (*Shared Word Count*) que se baseia na quantidade de palavras comuns entre documentos (Witten, 2004) e a Distância de Jaccard (Konchady, 2006) que representa o grau de união entre dois documentos.

2.1.3 Redução de Dimensionalidade

Um dos desafios constantes nas áreas de Recuperação da Informação e de Mineração de Textos é a alta dimensionalidade das matrizes geradas no pré-processamento, que passa a ser um fator crítico, em função do esforço computacional exigido. Por este motivo faz-se necessário a utilização de técnicas de redução de dimensões destas matrizes.

Na literatura duas formas de redução de dimensionalidade são mais citadas: Redução de Dimensionalidade por Seleção de Termos e Redução de Dimensionalidade por Extração de Termos.

Uma das técnicas (por Redução de Termos) bastante útil é a Indexação Semântica Latente. A Indexação Semântica Latente (do inglês *Latent Semantic Indexing*, LSI) tenta superar as deficiências da recuperação por combinação de termos, através de recursos estatísticos. Este método assume que há uma estrutura semântica oculta (latente) aos dados. Utiliza-se no LSI um modelo matemático para estimar esta estrutura latente, que liberta a discrepâncias constituídas pela polissemia[1] e pela sinonímia[2]

[1] Polissemia é a propriedade que uma mesma palavra tem de apresentar vários significados.

existente nos documentos. A descrição dos termos e dos documentos baseia-se na estrutura semântica latente dos termos. Entende-se por “estrutura semântica” a estrutura de correlação entre as palavras individuais que aparecem nos documentos. Semântico implica o fato de que os termos, em um documento, possam ser tomados como referentes ao documento ou ao assunto desse documento.

A principal vantagem do LSI em relação a outros métodos é que ele não depende das palavras de forma individual para indexar documentos, ou seja, o LSI utiliza conceitos ou tópicos para encontrar documentos relevantes.

O LSI tem como premissa que se dois documentos, A e B não possuem termos comuns entre si, mas ambos possuem termos comuns em relação a um terceiro documento C, então A e B devem ser considerados similares (Camargo, 2007).

O LSI mapeia os documentos em um vetor com menos termos no índice. Este índice passa a ser composto por conceitos em mais alto nível de abstração. A idéia é construir uma matriz de termos por documento na qual seus elementos representam a freqüência de cada termo no documento. A decomposição tem a propriedade de que os últimos fatores têm influência mínima sobre a matriz. Os fatores que menos influenciam podem ser descartados, diminuindo a dimensionalidade. Idealmente, a redução deve ser grande o suficiente para permitir que conceitos e detalhes relevantes da base de dados não sejam encobertos. De outra forma, não deve ser grande de forma a inviabilizar o processamento em função do esforço computacional.

O LSI utiliza a decomposição em valores singulares (do inglês SVD - *Singular Value Decomposition*). O SVD tem origem em álgebra linear e o resultado é a transformação da matriz de termos original em três outras matrizes. A multiplicação dessas matrizes reconstitui a matriz original.

O passo seguinte do LSI é a redução no tamanho da matriz de valores singulares, ou seja, reduzindo-a de um tamanho (r x r) para um tamanho (k x k), onde k é um número muito menor do que r. Todos os valores maiores que k serão transformados em

[2] Sinonímia é a relação que se estabelece entre duas ou mais palavras que possuem significados iguais

zero. Isto faz com que a matriz de valores singulares em passe a ter poucos elementos.

2.2 O PROCESSAMENTO

Nesta etapa a descrição dos dois principais de algoritmos utilizados na etapa de Processamento: Clusterização e Classificação.

2.2.1 Clusterização de Documentos

O objetivo da Clusterização é o agrupamento de documentos com características similares. Técnicas de clusterização são amplamente utilizadas em KDD. Genericamente os algoritmos de agrupamento podem ser divididos em dois tipos: os hierárquicos e os de aglomeração.

Nos algoritmos hierárquicos, os *clusters* são formados através de sucessivas divisões de elementos do grupamento original, gerando assim uma hierarquia – normalmente representada por uma árvore. Enquanto nos de aglomeração, inicialmente cada elemento é associado a um cluster distinto, e novos clusters vão sendo formados pela união dos clusters existentes.

Os algoritmos de clusterização já são bastante estudados nas áreas de estatística e matemática. Especificamente na área de *Data Mining* estes problemas têm a particularidade de estarem contextualizados em bases de dados de grande porte e com significativo número de variáveis.

Um dos principais objetivos da tarefa de clusterização de documentos é possibilitar a rápida localização de documentos em uma dada coleção. Quando uma coleção já está devidamente separada por clusters, encontrar um pequeno grupo de documentos relevantes torna-se uma tarefa fácil.

Para grandes coleções de documentos um algoritmo de clusterização pode vir a criar centenas de clusters em uma etapa inicial. Por esse motivo é necessário ter em mente alguns parâmetros para este trabalho, tais como: número mínimo e máximo de clusters, número mínimo de documentos para a formação de um cluster e o grau máximo de superposição entre clusters.

Algoritmo K-Means

Esta é uma técnica clássica de clusterização que pode ser adaptada para documentos. É amplamente utilizada para este fim e é razoavelmente eficiente - além de ser considerado um dos métodos mais simples. Este é um algoritmo de aprendizagem não supervisionada (e seu processo de separação) começa com a definição do usuário quanto ao número inicial k , de clusters.

Existem diversas maneiras de se criar um conjunto inicial de clusters neste método. Uma delas é através da seleção aleatória de documentos por cluster.

Outra forma de estabelecimento da seleção inicial de documentos por cluster é através do cálculo inicial de um vetor global que represente todos os documentos da coleção. Cada documento é então comparado com o vetor global e desta forma grupos iniciais (clusters) são criados a partir das similaridades de distância de cada documento a este vetor.

Para os dois tipos de formação original de clusters, o algoritmo calcula o centróide da cada grupo estabelecido. A partir daí cada documento será alocado no cluster que estiver mais próximo do centróide obtido na etapa anterior. O processo continua com o recálculo dos centróides a cada nova iteração, até que nenhum documento modifique sua posição, ou melhor, dizendo, mude de cluster.

Nota-se que desta maneira se obtém uma classificação que coloca cada ponto em apenas uma classe, portanto pode se dizer que este algoritmo faz uma classificação *hard* (hard clustering). Outros algoritmos trabalham com o conceito de classificação *soft* onde existe uma métrica que diz o quão ‘dentro’ de cada classe o ponto está.

O objetivo do algoritmo é minimizar a variância dos atributos dos pontos que estão dentro de um determinado segmento. Matematicamente podemos dizer que o k-means minimiza a função erro quadrático, onde existem k segmentos Si, com i = 1, 2, ... , k, e

μ_i i é o centróide, ou centro de um conjunto de pontos X_j, pertencentes S_i, conforme equação abaixo:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} |X_j - \mu_i|$$

Na Figura 2, um exemplo gráfico de como o algoritmo realiza a separação entre os grupos, conforme Figueiredo (2008):

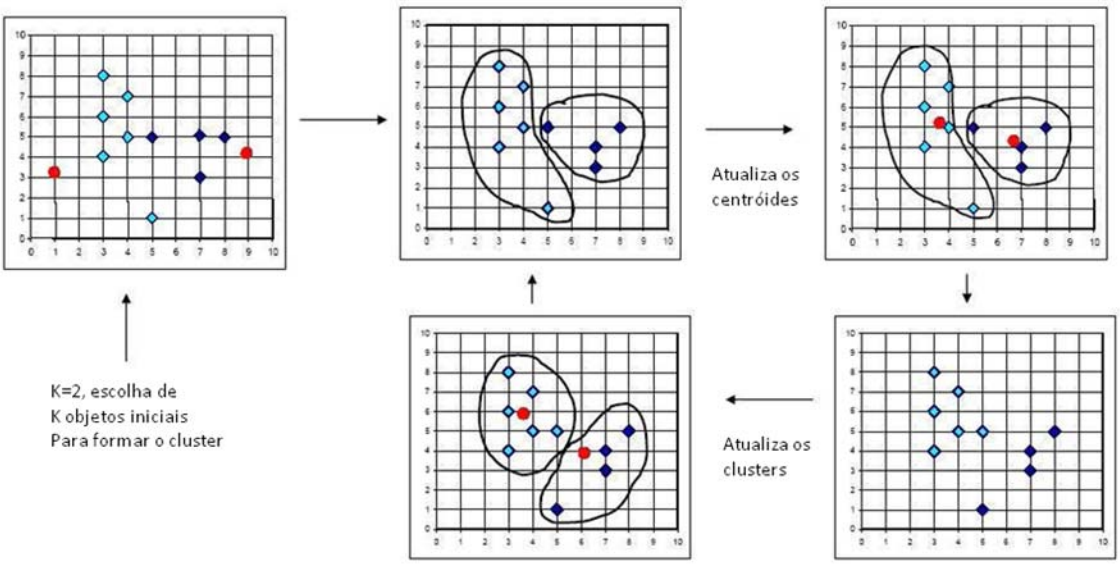


Figura 2: Exemplo gráfico de Clusterização utilizando o K-Means

Clusterização Hierárquica

Na clusterização hierárquica os grupos formados são representados em uma estrutura conhecida como *dendrograma*, que consiste em um tipo especial de árvore, na qual os nós pais agrupam os exemplos representados pelos nós filhos (Murtagh, 1983). Dessa maneira, um agrupamento hierárquico agrupa os dados de modo que se dois documentos são agrupados em algum nível, nos níveis mais acima eles continuam fazendo parte do mesmo grupo, construindo uma hierarquia de clusters. Essa técnica permite analisar os clusters em diferentes níveis de granularidade, pois cada nível do *dendrograma* descreve um conjunto diferente de agrupamentos.

Duas abordagens podem ser derivadas do clustering hierárquico: Aglomerativo (*Botton-up*) e Divisivo (*Top-down*). Na primeira abordagem, os dados são inicialmente distribuídos de modo que cada exemplo represente um cluster e, então esses clusters são recursivamente agrupados considerando alguma medida de similaridade, até que todos os exemplos pertençam a apenas um cluster

O Método "Ward" de Clusterização Hierárquica

Uma técnica multivariada que engloba diferentes algoritmos de classificação para organizar informações sobre variáveis e formar grupos homogêneos (Ward, 1963).

O Ward é um método que utiliza a análise da variância, para avaliar a distância entre os Clusters. É derivado de um processo hierárquico e aglomerativo. Este método tem por objetivo minimizar o quadrado da distância euclidiana às médias dos conglomerados. A distância euclidiana é a raiz quadrada da soma dos quadrados das diferenças de valores para cada variável.

No princípio, obtém-se-se m grupos; ou seja, um grupo para cada vetor componente da base de dados. Neste estágio inicial o erro interno é nulo para todos os grupos pois cada vetor que compõe cada grupo é o próprio vetor médio do grupo. Igualmente o desvio padrão para cada grupo é nulo. Na etapa subseqüente, cada possibilidade de aglutinação entre os grupos 2 a 2 é verificada, e é escolhido o agrupamento que causa o menor aumento no erro interno do grupo. São $m \times m$ verificações. Desta forma, para uma base de dados com m elevado, estas verificações exigem um grande esforço computacional caso o método seja implementado em computador. Nota-se que a cada iteração obtém-se $m - i$ grupos (i =número de iterações), no entanto, como o número de elementos pertencentes a cada grupo aumenta, é maior o número de cálculos para o erro interno de cada grupo."

De modo geral, este método tem um desempenho computacional mais lento, contudo, pode ser uma alternativa (de utilização) caso se utilize preliminarmente alguma técnica de redução de dimensionalidade da BOW (Bag of Words).

O Algoritmo de Expectation-Maximixation (EM)

EM é um método interativo de clusterização que converge para uma probabilidade aposteriori através da função de verossimilhança p(θ|x) (Figueiredo 2004):

A função de verossimilhança de *n* variáveis aleatórias X_1, X_2 , . . . ,Xn é definida como a densidade conjunta de *n* variáveis aleatórias, digamos fX_1. ...... X_n f$X_1$1,...,Xn (X_1, .. , Xn; Θ), considerada como função de Θ. Em particular, se X1, . . . ,Xn é uma amostra aleatória da densidade f(X;Θ), então a função de verossimilhança é *f*(X_1;Θ)*f*(X_2;Θ).....*f*(X_n;Θ).

O algoritmo de *Expectation-Maximization* estima os componentes de probabilidade presentes em um certo cluster. Em outras palavras, EM é um método de estimação de características de um determinado conjunto de dados, quando estes dados estão incompletos ou quando existem dados faltantes (Bilmes, 1998).

O EM é baseado na teoria Bayesiana que assume que o algoritmo vai estimar M clusters (ou classes) $C_{i,j}$ =1.......,M. Para cada um dos N vetores X_k, k=1,.....N, o algoritmo calcula a probabilidade $P(C_j/\ X_k)$ de pertencer a uma certa classe (Theodoridis and Koutroumbas 2003). O maior valor de probabilidade indica a classe do vetor.

O algoritmo atua de forma interativa em dois estágios: o estágio E (Expectation) e o estágio M (Maximization). Formalmente, $\hat{\theta}$(t) = {μ_j(t), Σ_j(t)}, j= 1, . . . M. O método tenta aproximar $\hat{\theta}$(t) da distribuição real quando t = 0, 1 .

O algoritmo pode ser dividido em duas partes principais:

- Etapa E: encontram-se os valores das estatísticas suficientes para os dados completos, com base nos dados incompletos e nas estimativas atuais dos parâmetros

- Etapa M: nesta etapa os parâmetros de estimação de máxima verossimilhança (ou seja, $\hat{\theta}$(t)) são atualizados.

Considerando que X =$\{X_1$,........., $X_m\}$ denote os dados observados em um conjunto de *m* instâncias ocorridas independentemente e seja Z = $\{Z_1$...... $Z_m\}$ os dados não observados destas mesmas instâncias e seja Y = X U Z, o total de dados.

E ainda: Z pode ser tratada como uma variável aleatória cuja distribuição de probabilidades depende do conjunto de parâmetros desconhecidos *Θ* e dos dados observados X. Analogamente, Y é uma variável aleatória porque é definida em termos da variável aleatória Z. Para descrever a forma geral do algoritmo EM, *h* denota a hipótese atual dos parâmetros *Θ,* e *h'* denota a hipótese revisada que é estimada em cada iteração de algoritmo EM.

Segundo Luna(2004), o algoritmo EM busca a hipótese *h'* de máxima verosssimilhança, isto é, que maximize E$[\ln P(Y|h')]$. Este valor esperado é calculado sob a distribuição de probabilidade de Y, que é determinada pelos parâmetros desconhecidos Θ. Sabendo-se que P $(Y|h')$ é a máxima verossimilhança de todos os dados Y, dada a hipóstese *h'*. Desta forma é desejável poder encontrar *h'* que maximize alguma função desta quantidade. Além disso, também é desejável maximizar o logaritmo desta quantidade $ln P(Y|h')$ que também maximiza $P(Y|h')$. E ainda (desejável), introduzir o valor esperado de $E[ln\ P(Y|h')]$ porque o total de dados Y é, ele próprio, uma variável aleatória.

Em geral não se sabe esta distribuição de Y, pois ela é determinada pelos parâmetros Θ que se está tentando estimar. Entretanto, o algoritmo EM usa sua hipótese atual *h*, no lugar do parâmetro Θ atual para estimar a distribuição de probabilidades de Y. Considerando a definição de uma função $Q(h'|\ h)$ que é $E[ln\ P(Y|h')]$ como uma função de *h',* sob a suposição que $\Theta = h$ e dada a porção observada X dos dados Y, ou seja:

$$Q(h'|\ h) = E[ln\ P(Y|h')|h, X]$$

Escreve-se esta função *Q* na forma *Q*(h'|h) para indicar que ela é definida em parte pela suposição que a hipótese atual *h* é igual a Θ.

O algoritmo em sua forma geral, repete os dois passos seguintes, até a convergência:

Passo E: calcula $Q(h'|\ h)$ usando a hipótese atual h e os dados observados X para estimar a distribuição de probabilidade sobre Y.

Passo M: troca a hipótese *h* pela hipótese *h'* que maximiza esta função Q

Como o EM é um método interativo, a sua utilização, principalmente para clusterização de grandes bases de dados, pode representar grande esforço computacional. Desta forma, uma alternativa é utilizar o K-Means como forma de estabelecer um *dataset* inicial de parâmetros θ. Neste trabalho esta alternativa (K-Means como parâmetro de inicialização) foi utilizada.

A seguir, na Figura 3, o *workflow* específico para o algoritmo de EM combinado ao K-Means, em referência a Thales (2008).

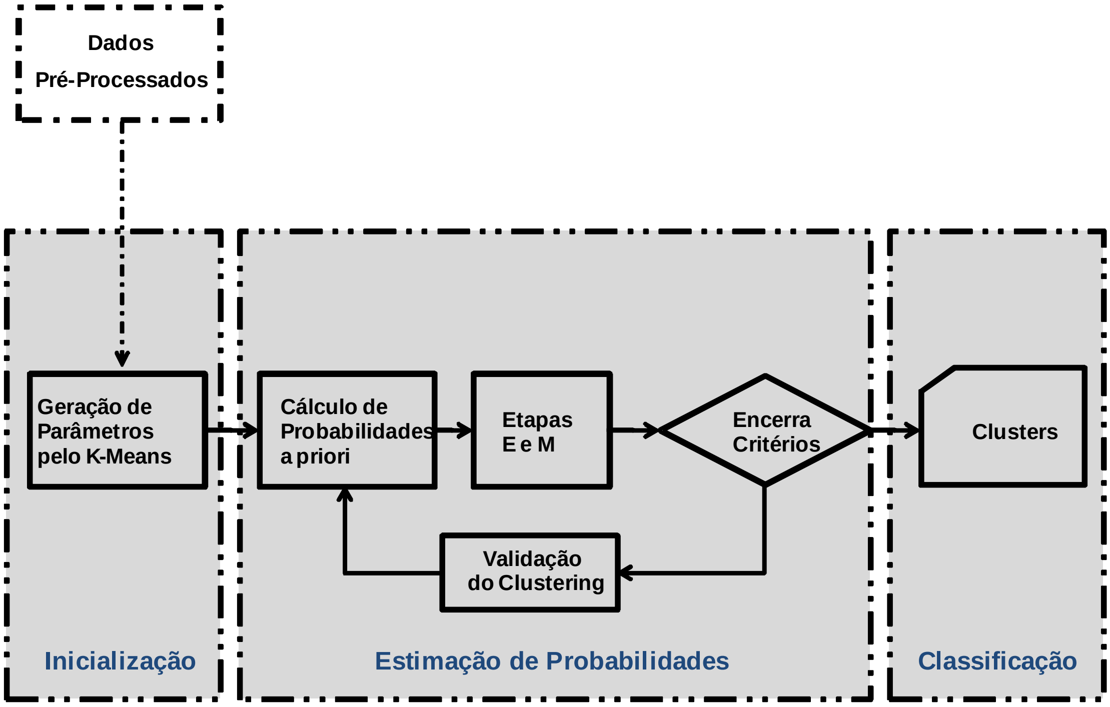


Figura 3: Workflow do Clustering por Expectation-Maximization

Inicialização no EM: utilizando um amostra, um set x é criado com os parâmetros de inicialização (*seed points*) do K-Means.

Estimação de Probabilidades no EM: este módulo realiza um procedimento interativo de sucessivas estimações de parâmetros e de validação do clustering. Este processo procura satisfazer ao número de classes estipuladas pelo usuário e também garante a maximização da distância entre clusters. Enquanto t cresce, um teste é realizado para checar se o algoritmo convergiu ou se um número máximo de interações foi atingido (conforme critério do usuário)

Classificação no EM: nesta etapa a classificação final é realizada. Cada um dos *n* documentos x_K é associado à classe com maior probabilidade, que é o mesmo que calcular P(C_j/x_k) > P(C_i /x_k); j≠ i e classificando x_k como C_j

2.2.2 Classificação de Documentos

Conceitos Básicos

Bancos de Dados são ricos em informações que podem ser usadas para tomadas de decisões inteligentes. Os processos de classificação e predição são duas formas de análise que podem ser utilizadas para extrair modelos que descrevam classes ou para predizer tendências. Este tipo de análise pode ajudar a prover um melhor entendimento de uma "massa de dados".

Um Classificador funciona como um processo em duas etapas. Na primeira etapa um Classificador é construído ao descrever um determinado subgrupo de classes ou conceitos. Esta é a chamada etapa de "aprendizado", onde utiliza-se um algoritmo para a construção do classificador baseando-se em uma "amostra para aprendizado" baseada em tuplas extraídas da base de dados, com suas respectivas classes associadas.

Desta forma, dada uma *tupla*, representada por um vetor n-dimensional X= {X_1, X_2, X_3,... X_N}; pode-se afirmar que cada uma destas (tuplas) pertence a uma classe C.

Um processo de classificação consiste basicamente em conseguir atribuir classes, baseando-se na combinação dos valores da tupla. Ou ainda, a combinação dos valores de uma tupla vai determinar, através do processamento do algoritmo que estiver sendo utilizado, a classificação desta tupla. Os algoritmos que realizam esta tarefa são chamados de modelos de Classificação ou simplesmente Classificadores.

A idéia é que uma vez obtido o modelo treinado, este possa fazer a classificação entre grupos de maneira análoga à classificação original do usuário e com a melhor acurácia possível. Sendo acurácia neste caso, definida por:

Acurácia = (Número de Classificações corretas) / (Número de classificações)

A seguir são descritos alguns métodos / algoritmos de classificação supervisionada, disponíveis na literatura.

O Classificador Bayesiano

O Classificador Baeysiano é baseado no Teorema de Bayes. Seja um determinado dado X cuja classe ainda seja desconhecida. Por sua vez, P(H|X) é a probabilidade a posteriori da hipótese H para o elemento X. De modo contrário, P(H) é a probabilidade apriori de H. Exemplo: P(H) é a possibilidade de qualquer fruta ser uma maçã, mesmo sem saber se esta fruta é vermelha ou arredondada.

Em resumo, o Teorema de Bayes é uma forma de calcular a probabilidade aposteriori P(H/X), a partir de P(H), P(X) e P(X/H), conforme a equação a seguir:

$$P(H|X) = P(X|H). P(H) / P(X)$$

Pode-se dizer então: a probabilidade aposteriori, P(H|X), é baseada em um maior volume de informações (ou com conhecimentos prévios) enquanto a probabilidade apriori, P(H), é independente de X.

De maneira similar, P(X|H) é a probabilidade de X condicionada em H. Voltando ao exemplo da maçã, P(X|H) é a probabilidade de X ser vermelho e arredondado, dado que se sabe que X é uma maçã.

Baseando-se no Teorema de Bayes descrito, o classificador Bayesiano é um classificador probabilístico e que pode ser bastante eficiente através de aprendizado supervisionado. Estudos comparativos de algoritmos de classificação tem encontrado no classificador Bayesiano performance equivalentes a métodos mais complexos como Redes Neurais e Arvores de Decisão (Han & Kamber, 2006). Por estes motivos, o Classificador Bayesiano é um dos mais utilizados na Classificação de Textos (McCallum & Nigam, 1998).

Support Vector Machine (SVM)

> O algoritmo de Classificação "Support Vector Machine" oferece uma bem fundamentada proposta de aprendizado supervisionado. Este tipo de

modelo apresenta bons resultados para problemas complexos de classificação. Contudo, este modelo é aplicável particularmente a problemas de classificação binária, como por exemplo: Consumidor Satisfeito / Consumidor Insatisfeito.

Em sua forma de utilização mais simples, o SVM se utiliza de um hiperplano que maximiza a margem de separação entre classes. Este hiperplano é definido por um subgrupo de exemplos, chamados de vetores de suporte, que fazem o trabalho de marcar, delimitar as fronteiras entre as classes produzidas pelo algoritmo.

A Figura 4 apresenta uma visão geométrica da construção do hiperplano ótimo para um espaço bidimensional, além da interpretação dos vetores-suporte.

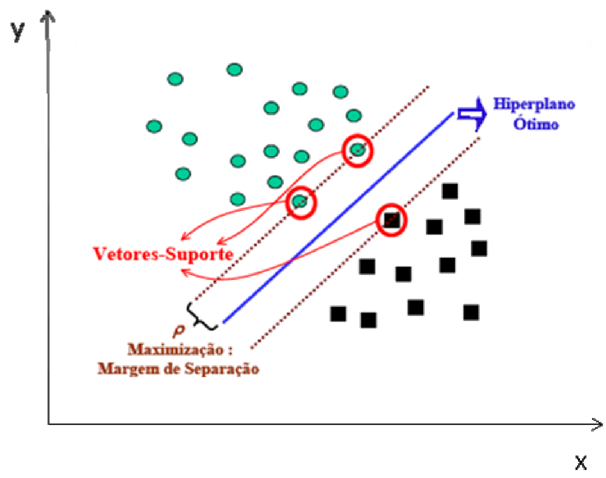


Figura 4: Representação do Hiperplano

Segundo Semolini (2002):

"Os vetores-suporte exercem um papel importante nas operações deste tipo de aprendizagem de máquina. Em termos conceituais, eles são os pontos que se encontram mais perto da superfície de decisão e, portanto, são os de classificação mais difícil. Como tal, eles têm uma relação direta com a localização da superfície de decisão."

Em comparação com 16 outros métodos de classificação, o SVM obteve boa performance. Para classificação, uma das vantagens percebidas foi uma redução do esforço computacional em relação aos demais métodos (Mayera, 2003)

2.3 OPINION MINING

É um conjunto de metodologias que complementam a forma comum de processamento textual, ao inserir a análise de frases, palavras ou expressões que de alguma forma expressem sentimentos positivos ou negativos sobre um produto ou serviço. Segundo Liu B.(2006), dado um conjunto de textos, onde são relatadas as opiniões de clientes, é possível classificar cada documento de uma dada coleção em duas classes: classe de sentimentos positivos e classe de sentimentos negativos.

Uma das motivações para o desenvolvimento desta área (Opinion Mining) é a Internet, que vem mudando drasticamente a forma como as pessoas expressam suas opiniões. Os consumidores atualmente podem a qualquer momento opinar em fóruns, blogs, grupos de discussão, ou mesmo sites especializados em classificação de produtos. Melhor dizendo: os consumidores atualmente podem e efetivamente geram informação e mídia espontânea relevante e de caráter pessoal sobre os produtos e serviços que consomem.

2.3.1 Classificação do Tipo de Sentimento

É o tipo de situação em que Opinion Mining é um problema de classificação de textos. Dado um determinado dataset D de textos, uma Classificação de Sentimentos é aquela que classifica cada um dos documentos d desta coleção em duas classes, positiva e negativa. Este primeiro filtro de classificação nos textos é feito pela exploração da existência de palavras com conteúdo positivo, negativo e neutro.

A maior parte de trabalhos pregressos nesta área estava focada na distinção binária, ou seja, de textos com conotação positiva ou conotação negativa. Mas é especialmente interessante poder estabelecer um "ranking" a gradação de intensidade destas inferências de opinião. Koppel & Scler (2005) mostram ainda que é muito importante utilizar exemplos com neutralidade em relação às polaridades (positiva ou negativa), pois isto melhora o processo de classificação, além de permitir uma terceira classificação, ou seja, neutra.

Uma possibilidade adicional nesta etapa do processamento é a classificação baseada em sentenças - ou seja, a classificação entre sentenças positivas e negativas. Para este caso em específico, Turney(2002) assim como Santorini(1990) trouxeram contribuições nesta área através do algoritmo denominado de "Part-Of – Speech (POS) Tagging". A seguir uma descrição deste algoritmo em específico.

2.3.2 POS: Part of Speech Tagging (POS)

Nesta metodologia utilizam-se como informações relevantes as características sintáticas ou morfológicas das palavras e as suas respectivas inter-relações. Entendam-se características sintáticas como a atribuição da classificação de substantivos, adjetivos, verbos, conjunções e interjeições, como parte da informação que é utilizada na técnica.

O algoritmo de utilização do POS foi demonstrado por Turney (2002) e consiste em três etapas:

Etapa 1: dado que estas funções sintáticas (adjetivos e advérbios) costumam ser bons indicativos de conceitos subjetivos, a análise de adjetivos ou advérbios, vistos isoladamente, pode levar a ambigüidades. Determinados adjetivos combinados a determinados substantivos, podem indicar uma característica (ou opinião) positiva, enquanto o mesmo adjetivo combinado a um substantivo diferente pode denotar uma característica (ou opinião) negativa. Por este motivo o algoritmo demonstrado por Turney extrai sempre duas palavras consecutivamente de forma a confrontar a alguma das opções que serão descritas na Tabela 2.

Melhor explicando: dada uma lista-padrão de binômios de palavras, o primeiro passo do algoritmo de "POS" é identificar que tipo de seqüência ocorre; ou seja, dada uma lista padronizada de seqüências, o algoritmo identifica o binômio como uma das possibilidades desta lista[3]. Conforme Turney (2002), a extração dos binômios deve seguir o padrão abaixo:

	Primeira Palavra	Segunda Palavra	Terceira Palavra (não extraída)
1.	JJ	NN ou NNS	nenhuma
2.	RB, RBR ou RBS	JJ	nem NN nem NNS
3.	JJ	JJ	nem NN nem NNS
4.	NN ou NNS	JJ	nem NN nem NNS
5.	RB, RBR ou RBS	VB, VBD, VBN ou VBG	nenhuma

Tabela 2: Tags para Extração de Binômio de Palavras

Sendo:

JJ = Adjetivo

NN = Nome no singular

NNS = Nome no Plural

RB = Advérbio

RBR = Advérbio Comparativo

RBS = Advérbio Superlativo

VB = Verbo

VBD = Verbo no Passado (Past Tense)

VBN = Verbo no Passado (Past Participle)

VBG = Verbo no Gerúndio ou no Presente

[3] Obviamente como o algoritmo fora desenvolvido para a língua inglesa – alguns dos tipos de binômios originalmente propostos por Turney(2002) podem não fazer sentido para a língua portuguesa.

Para exemplificar a separação de uma sentença com o padrão estabelecido por Turney (2002) segue a frase: “essa câmera produz bonitas fotos”. Neste caso o binômio “bonitas fotos” satisfaz a condição 1 na lista apontada na Tabela 2, pelo fato de ser um adjetivo seguido de um substantivo.

Etapa 2: tendo a etapa de marcação (*Tagging*) sido realizada, a próxima etapa do método é o cálculo da Orientação Semântica (OS) de cada sentença analisada. Este é outro indicador no processo de Classificação de Sentimentos das frases extraídas.

A forma de cálculo da OS é feito utilizando-se o cálculo de “Pointwise Mutual Information” (PMI), que é o cálculo matemático referente à ocorrência de dois eventos, no caso, das duas palavras analisadas na primeira etapa do método, em um mesmo dataset, sendo:

$$PMI = \log\left(\frac{P(X,Y)}{P(X).P(Y)}\right)$$

Onde:

- $P(X,Y)$ é a probabilidade de co-ocorrência dos termos x e y em uma mesma sentença

- P(X).P(Y) é a probabilidade de co-ocorrência destes dois termos, considerando que eles sejam estatisticamente independentes.

Portanto, PMI é o grau de dependência estatística entre $termo_x$ e o $termo_y$. Ou ainda: PMI é a medida do volume de informação adquirida na presença de uma das palavras analisadas, quando a outra palavra é observada.

Etapa 3: Cálculo da Orientação Semântica

Orientação Semântica é uma medida de inferência do quanto uma expressão se aproxima de uma expressão de referência positiva ou de referência negativa. Turney (2002) utilizou as palavras “excellent” e “poor” como referência positiva e negativa, respectivamente.

SO (frase) = PMI(frase, “excellent”) – PMI(frase, “poor”)

Turney calculou essas probabilidades realizando “*queries”* e coletando o número de “*hits”* que a ferramenta de busca retornava. Desta forma, o PMI pode ser calculado utilizando-se recursos de Recuperação da Informação. A frequência de ocorrência de termos em relação a palavras (negativas ou positivas) pode ser feita através do uso de *queries* do tipo “termo a ser pesquisado” NEAR “palavra de caráter positivo”, para a verificação da conotação postiva. Faz-se o mesmo para a verificação em relação a uma palavra de caráter negativo. Dentre os “motores de busca” disponíveis na web, a que oferece este recurso de maneira mais simples é o Altavista (www.altavista.com), que através do Operador Lógico (NEAR), identifica a quantidade de links que possuam frases onde o termo pesquisado aparece a uma distância de até dez termos da palavra que venha imediatamente após o operador lógico NEAR.

Etapa 4: é o cálculo da Orientação Semântica para cada expressão que possua correspondência na Tabela 2.

Nesta etapa, cada texto tem seu cálculo de Orientação Semântica efetuado através da soma dos PMIs de cada expressão de palavras consecutivas que se adequem à adaptação da Tabela 2. A soma de todos os PMIs dá o valor da Orientação Semântica do texto ou do dataset em questão. A seguir, nas tabelas 3 e 4, um exemplo de aplicação do cálculo de Orientação Semântica (Turney, 2002):

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732
Average Semantic Orientation		0.322

Tabela 3: Exemplo de aplicação de Orientação Semântica

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
little difference	JJ NN	-1.615
clever tricks	JJ NNS	-0.040
programs such	NNS JJ	0.117
possible moment	JJ NN	-0.668
unethical practices	JJ NNS	-8.484
low funds	JJ NNS	-6.843
old man	JJ NN	-2.566
other problems	JJ NNS	-2.748
probably wondering	RB VBG	-1.830
virtual monopoly	JJ NN	-2.050
other bank	JJ NN	-0.850
extra day	JJ NN	-0.286
direct deposits	JJ NNS	5.771
online web	JJ NN	1.936
cool thing	JJ NN	0.395
very handy	RB JJ	1.349
lesser evil	RBR JJ	-2.288
Average Semantic Orientation		-1.218

Tabela 4: Exemplo de aplicação de Orientação Semântica

O exemplo acima indicou um caso de Orientação Semântica (OS) negativa, uma vez que a média das Orientações Semânticas foi negativa.

2.3.3 Classificação de Opinião utilizando Métodos Convencionais

A forma mais simples de abordar o problema de avaliação das opiniões embutidas em sentenças e, consequentemente, em conjuntos de textos, é a utilização de métodos convencionais de classificação, como por exemplo Naive Bayes ou SVM.

Este tipo de abordagem foi experimentada por Pang et al (2002), utilizando textos referentes a comentários de consumidores sobre filmes, clasificados por estes em positivos e negativos. As taxas de acertos foram de 81% e 82,9% para os métodos de Naive Bayes e SVM, respectivamente.

Sendo um método supervisionado, necessita obviamente de conjuntos pré-classificados para possibilitar a aplicação destes métodos.

2.4 RESUMO DO CAPÍTULO

Neste capítulo foram abordados os processos para a transformação de informação não estruturada, sob a forma de texto, em informação relevante, através da limpeza de dados e as subseqüentes etapas de pré-processamento. Além disso, foram vistas em detalhes as etapas e particularidades do processamento da base já devidamente tratada e com suas dimensões reduzidas.

É importante ressaltar a enorme quantidade de informação não estruturada sob a forma de arquivos eletrônicos nos dias de hoje, abrindo desta forma um grande horizonte para o desenvolvimento de pesquisas e aplicações nesta área do conhecimento. Além disso, possibilidades ainda maiores se abrem quando tais técnicas (de Mineração de Textos) são combinadas a técnicas convencionais de Data Mining ou às técnicas ainda não utilizadas de maneira combinada como o Georeferenciamento e a Geoestatística.

O capítulo seguinte (Capítulo 3) aborda as técnicas e processos relacionados a Sistemas de Informação Geográfica para na sequência na descrição da pesquisa propriamente dita (Capítulo4) abordar os dois assuntos de maneira integrada (Mineração de Textos e Georeferenciamento).

CAPÍTULO 3: SISTEMAS DE INFORMAÇÃO GEOGRÁFICA

Esta etapa do trabalho apresenta conceitos de geoprocessamento e modelagem geoestatística, buscando estabelecer um método de visualização e análise espacial das características da opinião de consumidores, obtidas através da Mineração de Textos.

Geoprocessamento é a área do conhecimento para o tratamento de informação geográfica e que vem ao longo dos últimos anos sendo aplicado em diversas áreas, como por exemplo: Cartografia, Recursos Naturais, Energia e Planejamento.

Ferramentas de Geoprocessamento também conhecidas como Sistemas de Informação Geográfica (SIG's) utilizam Informações relacionadas a uma Base de Dados que contém coordenadas (latitude, longitude) e que, por este motivo, possibilitam a realização de análises abrangentes, por permitirem uma visão espacial do problema analisado.

Segundo Camara et al (2004), os SIG's devem tentar responder a perguntas relacionadas ao "porque" ao "quando" e "onde", ou seja, descrevendo o fenômeno quantitativamente e prevendo sua evolução espacial e temporal.

3.1 CONCEITOS BÁSICOS DE ESPAÇO GEOGRÁFICO E DE INFORMAÇÃO ESPACIAL

O espaço geográfico tem como característica a dualidade da localização geográfica (expressa como coordenadas em um espaço geográfico) e atributos descritivos (que podem ser representados num banco de dados convencional). De forma intuitiva, pode-se definir o termo "espaço geográfico" como uma coleção de localizações na superfície da Terra, sobre a qual ocorrem os fenômenos geográficos. O espaço geográfico define-se, portanto, em função de suas coordenadas, sua altitude e sua posição relativa. Sendo um espaço localizável, o espaço geográfico é possível de ser cartografado (Dolfus, 1991).

A noção de informação espacial está relacionada à existência de objetos com propriedades, que incluem sua localização no espaço e sua relação com outros objetos. Estas relações incluem conceitos topológicos (vizinhança, pertinência), métricos (distância) e direcionais ("ao norte de", "acima de"). Deste modo, os conceitos de espaço geográfico e informação espacial são duas formas complementares de conceituar o objeto de estudo do Geoprocessamento. Estas formas irão levar à dualidade conceitual na modelagem espacial, onde a noção absoluta de espaço geográfico leva à idéia de conjuntos de campos geográficos e a noção relativa de informação espacial conduz à postulação da existência de conjuntos de objetos georeferenciados.

3.1.1 Relações Espaciais nos Fenômenos Geográficos

Os diferentes fenômenos geográficos, ao se distribuírem sobre a superfície da Terra, estabelecem padrões de ocupação. Ao representar tais fenômenos, o Geoprocessamento procura determinar e esquematizar os mecanismos implícitos e explícitos de inter-relação entre eles. Estes padrões de inter-relação podem assumir diferentes formas:

- correlação espacial: um fenômeno espacial está relacionado com o entorno de forma tão mais intensa, quanto maior for a proximidade de localização. Diz-se informalmente que "*coisas próximas são parecidas*";

- correlação temática: as características de uma região geográfica são moldadas por um conjunto de fatores. Assim, o clima, as formações geológicas, o relevo, o solo, a vegetação formam uma totalidade inter-relacionada. Deste modo, podem-se traçar pontos de correspondência entre o relevo e o solo ou o solo e a vegetação de uma região;

- correlação temporal: a fisionomia da Terra está em constante transformação, em ciclos variáveis para cada fenômeno. Cada paisagem ostenta as marcas de um passado mais ou menos remoto, apagado ou modificado de maneira desigual, mas sempre presente (Dolfus, 1991);

- correlação topológica: de particular importância na representação computacional, as relações topológicas como adjacência, pertinência e intersecção, permitem estabelecer os relacionamentos entre os objetos geográficos que são invariantes à rotação, à translação e à escala.

3.2 MODELOS DE DADOS PARA GEOPROCESSAMENTO

O processo de Modelagem com Geoprocessamento é a forma que se dispõe para traduzir o mundo real em outros domínios. Uma das abordagens mais úteis para este problema é o chamado "paradigma dos quatro universos" (Gomes e Velho, 1995), que distingue:

- universo do mundo real: que inclui as entidades da realidade a serem modeladas no sistema;

- universo matemático (conceitual): que inclui uma definição matemática (formal) das entidades a serem incluídas no modelo;

- universo de representação: onde as diversas entidades formais são mapeadas para representações geométricas;

- universo de implementação: onde as estruturas de dados e algoritmos são escolhidos, baseados em considerações como desempenho, capacidade do equipamento e tamanho da massa de dados. A seguir (Figura 5) uma representação destes Universos:

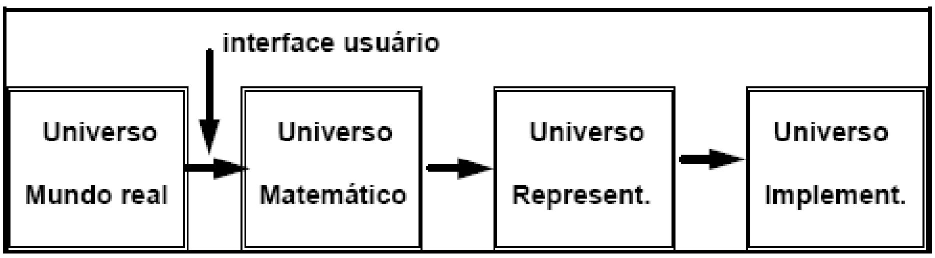


Figura 5: os quatro universos segundo Monteiro e Câmara (2007)

3.3 MODELAGEM ESPACIAL:

Modelos Estatísticos são constituídos de pelo menos três elementos: variáveis, relacionamentos e processos. Dependendo do objetivo específico a que o modelo se destina, pode-se dar maior ênfase a um ou outro aspecto destes que foram mencionados. Além disso, podem-se classificar os modelos em empíricos e de sistemas, a seguir:

- Modelos Empíricos: avaliam os relacionamentos entre as variáveis a partir da premissa de que as correlações existentes e analisadas no passado continuarão a existir no futuro. ME's são caracterizados pela simplicidade das variáveis envolvidas.

- Modelos de Sistemas: são descrições matemáticas de processos complexos, cujas variáveis interagem entre si. Este tipo de modelo atém-se principalmente às interações entre todos os componentes de um sistema (Lambin, 1994). Modelos de Sistemas procuram descrever o sistema como um todo, isto é, tentam representar as interações entre todos os seus componentes. Segundo Câmara et al (2004), uma característica chave destes modelos é a eficiência com que abordam as visões espaciais, implementando conceitos de relação de vizinhança e suportando a utilização de múltiplas escalas. Os Modelos de Georeferenciamento / Geoestatística são portanto uma particularidade dos Modelos de Sistemas.

3.3.1 Conceitos Básicos de Modelagem Geoestatística

Os métodos geoestatísticos (ou simplesmente geoestatística) foram desenvolvidos graças aos estudos do engenheiro de minas Georges Matheron, na França, no final da década de 50 e início da década de 60. Estes métodos estão fundamentados na Teoria das Variáveis Regionalizadas, que foi formalizada por Matheron (1965), a partir de estudos práticos desenvolvidos por Daniel G. Krige no cálculo de reservas nas minas de ouro na África do Sul.

A geoestatística é a aplicação destes métodos matemáticos e estatísticos a problemas das Ciências da Terra, com o objetivo principal de estimar simultaneamente um conjunto de variáveis espacialmente correlacionadas (variáveis regionalizadas). A utilização desta técnica é indicada quando uma das variáveis não foi amostrada em quantidade suficiente devido a dificuldades experimentais ou altos custos, proporcionando estimativas de precisão aceitável.

Assim, os métodos geoestatísticos fornecem um conjunto de ferramentas para entender a aleatoriedade dos dados, mas com possível estruturação espacial, estabelecendo, desse modo, uma relação espacial.

Teoria das variáveis regionalizadas

Na teoria das variáveis regionalizadas, Z(x) pode ser definida como uma variável aleatória que assume diferentes valores Z em função da posição x dentro de certa região S e representa pares de coordenadas (X_i,Y_i), conforme Figura 6. O ponto de referência para o sistema de coordenadas é arbitrário e fixado a critério do interessado (Vieira, 2000). O conjunto de variáveis Z(x) medidas em toda a área S pode ser considerada uma função aleatória Z(x) uma vez que, segundo Isaaks e Srivastava (1989), são variáveis aleatórias, regionalizadas e assume-se que a dependência entre elas é especificada por algum mecanismo probabilístico.

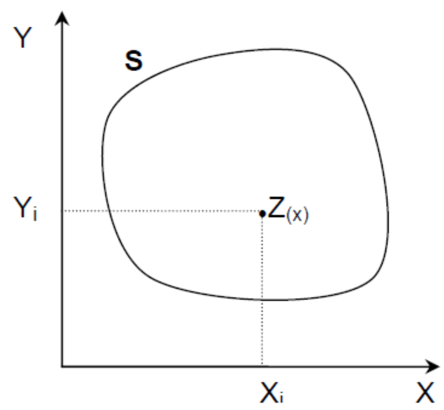


Figura 6: Variável Aleatória Regionalizada Z(x)

Relação Espacial

Relações espaciais estão presentes, tanto em linguagens de consulta espacial, quanto em aplicações geográficas. Segundo Papadias & Theodoridis (1997), a representação e o processamento das relações espaciais são cruciais nas aplicações geográficas porque, frequentemente, no contexto do espaço geográfico, relações entre entidades espaciais são tão importantes quanto as próprias entidades. Dependendo do domínio de aplicação, algumas relações espaciais se tornam mais importantes que outras.

Segundo Egenhoger (1991), as relações espaciais são agrupadas em três categorias:

a) topológicas: são aquelas que descrevem os conceitos de vizinhança, incidência, sobreposição, mantendo-se invariante ante as transformações como escala e rotação;

b) métricas: são consideradas em termos de direções e distâncias, sendo que as relações direcionais são aquelas que descrevem a orientação no espaço como, por exemplo, "norte" e "sul", e as relações de distâncias são aquelas que dependem de definições métricas no sentido de se "parametrizar" o quanto é perto ou longe como, por exemplo, "perto de" e "longe de"; tal parametrização dependerá das circunstâncias e das entidades geográficas relacionadas;

c) de ordem: são aquelas que expressam a ordem, total ou parcial, dos objetos espaciais,sendo descritas por preposições como "em frente a", "atrás de", "acima de" e "abaixo de".

Semivariograma

A estimativa da dependência entre amostras vizinhas no espaço pode ser realizada através da autocorrelação[4] que é de grande utilidade quando se está fazendo amostragem em uma direção. Quando a amostragem envolve duas direções (x,y) o instrumento mais indicado na estimativa da dependência entre amostras é o semivariograma (Silva, 1988).

[4] Em estatística autocorrelação é uma medida que informa o quanto o valor de uma de uma variável aleatória é capaz de influenciar seus vizinhos. Por exemplo, o quanto a existência de um valor mais alto condiciona valores também altos de seus vizinhos.

O semivariograma analisa o grau de dependência espacial entre amostras dentro de um campo experimental, além de definir parâmetros necessários para a estimativa de valores para locais não amostrados, através da técnica de krigagem (Salviano, 1996).

O semivariograma é a ferramenta básica, que permite descrever quantitativamente a variação no espaço de um fenômeno regionalizado (Huijbregts, 1975). A natureza estrutural de um conjunto de dados (assumido pela variável regionalizada) é definida a partir da comparação de valores tomados simultaneamente em dois pontos, segundo uma determinada direção. A função de semivariância γ(h) é definida como sendo a esperança matemática do quadrado da diferença entre os valores de pontos no espaço, separados por uma distância h, conforme a seguinte equação:

$$y(h) = (1|2)\ \mathrm{E}\left\{[Z_{(x)} - Z_{(x+h)}]^2\right\}$$

E pode ser estimada por:

$$y^*(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(X_i) - Z(X_i + h)]^2$$

em que N(h) é o número de pares de valores medidos Z(xi), Z(xi+h), separados por um vetor h (Journel & Huibregts, 1978).

O gráfico de λ(h) versus h representa o semivariograma, que permite obter a estimativa do valor de semivariância para as diferentes combinações de pares de pontos e assim analisar o grau de dependência espacial da variável estudada e definir os parâmetros necessários para a estimativa de suas características em locais não amostrados (Souza, 1999).

A medida que h aumenta λ(h) também aumenta pois é de se esperar que amostras tiradas a uma pequena distância entre si apresentem $(Z_{(x)} - Z_{(x+h)})^2$ menores que aquelas obtidas a distâncias maiores (Camargo,1997). O ajuste do modelo matemático aos dados no gráfico, ou seja, a uma função, define os parâmetros do semivariograma, que são: efeito pepita (Co), que é o valor de quando h=0; quando h aumenta freqüentemente, aumenta até uma distância a, chamada de alcance (a) da dependência espacial; e a partir da qual (h) neste ponto é chamado de patamar (C+Co), cujo valor é aproximadamente igual à variância dos dados, se ela existe, e é obtido pela soma do efeito pepita e a variância estrutural (C) conforme pode ser observado na Figura 7.

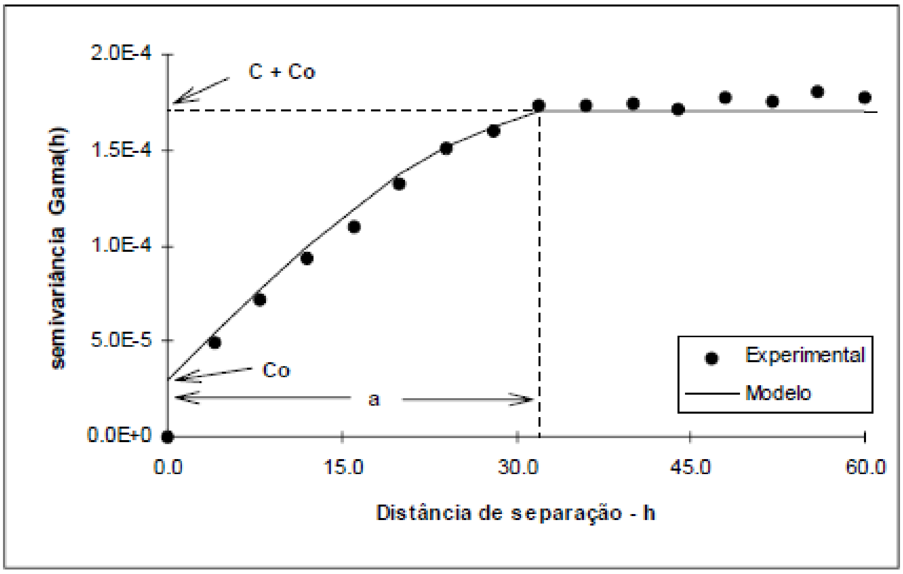


Figura 7: Semivariograma Ajustado

Amostras separadas por distâncias menores do que o alcance são espacialmente dependentes, enquanto aquelas separadas por distâncias maiores, não são, ou seja, um semivariograma igual à variância dos dados implica em variação aleatória. O alcance também é utilizado para definir o raio de ação ("range") máximo de interpolação por krigagem, onde os pesos utilizados na ponderação podem afetar os valores estimados (Souza, 1992).

No comportamento típico de um semivariograma ajustado, o valor de semivariância aumenta à medida que aumenta a distância de separação entre os pontos, até estabilizar-se, ou seja, atingir um patamar (Vieira, 2000). O patamar ("sill") é atingido quando a variância dos dados se torna constante com a distância entre as amostras. O valor de (h) nesse ponto é aproximadamente igual a variância total dos dados – este é um parâmetro importante, pois permite a determinação da distância limite entre dependência e independência entre das amostras (Silva, 1988).

O efeito pepita, que é um parâmetro importante do semivariograma, reflete o erro analítico, indicando uma variabilidade não explicada (ao acaso) de um ponto para o outro, que pode ser devida tanto a erros de medidas ou microvariação não detectada em função da distância de amostragem utilizada, sendo impossível quantificar a contribuição individual dos erros de medições ou da variabilidade.

Dependendo do comportamento do semivariograma para grandes valores de h, o modelo a ser usado pode ser classificado em duas categorias: a) modelo sem patamar e b) modelo com patamar.

Os modelos com patamar normalmente são ajustes que representam a estacionariedade de segunda ordem, onde a semivariância aumenta com o aumento da distância entre amostras, até atingir o patamar, onde se estabiliza (Machado, 1994). Já os modelos sem patamar satisfazem apenas a hipótese intrínseca e os semivariogramas podem ser definidos, mas não se estabilizam em nenhum patamar.

Krigagem

É um método de regressão usado em geoestatística para aproximar ou interpolar dados. A teoria de Kriging foi desenvolvida a partir dos trabalhos do seu inventor, Daniel G. Krige e pelo matemático francês Georges Matheron, no começo dos anos sessenta. Na comunidade estatística, também é conhecido como "Processo Gaussiano de Regressão".

Conhecido o semivariograma da variável, e havendo dependência espacial entre as amostras, podem-se interpolar valores em qualquer posição no campo de estudo, sem tendência e com variância mínima (Vieira, 2000). Este método de interpolação chama-se Krigagem e tem como base os dados amostrais da variável regionalizada e as propriedades estruturais do semivariograma obtido a partir destes dados; o que permite visualizar o comportamento da variável na região através de um mapa de "isolinhas" ou de superfície. Segundo Salviano (1996) o valor estimado da variável é dado pela expressão:

$$Z^*(x_0) = \lambda_0 + \sum_{i=1}^{n} \lambda_i \, Z(X_i)$$

na qual n é o número de vizinhos medidos, $Z(X_i)$ utilizados na estimativa da variável

e *i* são os ponderadores aplicados a cada $Z(X_i)$, os quais são selecionados de forma que a

estimativa obtida seja não tendenciosa. Isto implica em assumir duas condições:

$$E\{Z^*(X_0) - Z(X_0)\}=0$$

$$Var\ \{Z^*(X_0) - Z(X_0)\} = E\{[Z^*(X_0) - Z(X_0)]^2\} = \text{mínima}$$

Estas duas condições garantem que o estimador de krigagem seja o melhor estimador linear não tendencioso (BLUE = *Best Linear Unbiased Estimator*) pois apresenta variância mínima e não tendencioso por assegurar que o somatório dos pesos é igual à unidade.

3.4 SOFTWARES PARA ANÁLISE ESPACIAL

A seguir um resumo dos principais softwares de Análise Espacial disponíveis no mercado:

GSTAT – Software para Geoestatística

- descrição: ambiente para desenvolvimento de programas em geoestatística

- disponibilidade: software livre em www.gstat.org

- funções: análise Exploratória, estatísticas descritivas, cálculo de variograma (2D e 3D). Estimação: krigagem simples, ordinária e universal (com modelo de tendência), co-krigagem, krigagem por indicação, simulação seqüencial (gaussiana e por indicação), com suporte a variáveis contínuas ou categóricas.

ClusterSeer – Clustering de Processos Pontuais

- descrição: Programa para detecção de clusters (conglomerados) associados a eventos

- disponibilidade: Software comercial em www.terraseer.com

- funções: Detecção de Conglomerados Espaciais: testes focados (Diggle, Bithell, Besag e Newell, Turnbull) e globais (Besag e Newell, função K de Ripley). Detecção de Conglomerados Espaço-Temporais (Kulldorff)

SpaceStat – Análise Espacial de Áreas

- descrição: Software para análise espacial de áreas, com ênfase em técnicas de regressão espacial.

- contato comercial em www.spacestat.com

- funções Análise Exploratória._estatísticas descritivas, índice I de Moran (global e local), mapa de Moran, índice C de Geary, com testes de hipóteses sobre autocorrelação espacial. Regressão por mínimos quadrados e regressão espacial.

Spring

- descrição Software de geoprocessamento de propósito geral, com funções de processamento de imagens, modelagem de terreno, álgebra de mapas e consulta a bancos de dados geográficos.

- disponibilidade Software livre em www.dpi.inpe.br/spring

- funções de Análise Espacial Análise Exploratória: estatísticas descritivas, cálculo de variograma (2D e 3D), índice I de Moran (global e local), mapa de Moran, índice C de

Geary, com testes de hipóteses sobre autocorrelação espacial. Detecção de conglomerados: função K de Ripley, vizinho mais próximo e índices locais de Moran. Estimador de densidade: "*kernel estimator*". Estimação: krigeagem simples e ordinária, krigeagem por indicação, simulação seqüencial (gaussiana e por indicação), com suporte a variáveis contínuas ou categóricas.

TerraView

- descrição: software de Análise e Visualização de Dados baseado na Biblioteca Terralib.

- disponibilidade: software livre em www.terralib.org

- funções de Análise Espacial, Análise Exploratória: estatísticas descritivas, cálculo de variograma (2D e 3D), índice I de Moran (global e local), mapa de Moran, índice C de Geary, com testes de hipóteses sobre autocorrelação espacial.

ArcGIS:Geostatistical Analyst

- descrição Extensão do ArcGIS (software de geoprocessamento de propósito geral)

- contato comercial em: www.esri.com

- funções de Análise Espacial e Análise Exploratória: estatísticas descritivas, cálculo de variograma (2D e 3D), análise de tendências. Estimação: krigagem simples e ordinária, krigagem por indicação, co-krigagem e krigagem disjuntiva

MapInfo Professional

- descrição: Sotware de Geo-Referenciamento de uso empresarial

-disponibilidade: Comercial em www.mapinfo.com ou em www.geograph.com.br (representante comercial no Brasil)

- funções: Associação automática de dados a mapas através da função da geocodificação, criação de mapas temáticos. Capacidade de sobreposição de mapas a arquivos *raster* provenientes de: mapas em papel, fotos aéreas e imagens de satélites Importação e exportação de arquivos em formatos gráficos como, por exemplo, o DXF do AutoCad. Capacidade de alterar a projeção de seu mapa para exibição ou digitalização. Integração de critérios geográficos a buscas em banco de dados.

Google Maps / Google Earth

Uma possibilidade adicional de representação de informações georeferenciadas é a utilização dos recursos do Google Maps e Google Earth. Ambos disponibilizam mapas de todos os continentes com informações de relevo, de ruas e ainda de pontos de interesse comercial, além de informações postadas pelos usuários de todo o mundo.

Do ponto de vista de um processo de georeferenciamento com utilização acadêmica e/ou profissional, algo bastante interessante (que já é possível) é a utilização de uma aplicação com código aberto, para a criação de pontos e polígonos em um conjunto de camadas, correspondentes ao espaço geográfico, disponibilizados pela Google. Isto é feito através da linguagem KML.

KML é um padrão aberto oficialmente conhecido como OpenGIS® KML Encoding Standard (OGC KML). Ele é mantido pela *Open Geospatial Consortium, Inc*. (OGC). A especificação completa do OGC/KML pode ser encontrada em http://www.opengeospatial.org/standards/kml/.

O KML utiliza uma estrutura de "*tags*" com elementos e atributos aninhados e se baseia no padrão XML. Todas as *tags* diferenciam maiúsculas e minúsculas e devem aparecer exatamente como listadas na Referência do KML. A referência indica as "*tags*" que são opcionais. Em determinado elemento, as tags devem aparecer na ordem mostrada na referência. A utilização do KML em mapas do Google Maps / Google Earth é evidenciada na etapa de Metodologia de Pesquisa e Estudo de Caso deste trabalho.

3.7 RESUMO DO CAPÍTULO

Neste capítulo foi possível relacionar as principais características de Sistema de Informação Geográfica, de Georeferenciamento e de Geoestatística, assim como as técnicas que são utilizadas para identificação de correlações espaciais. Além disso, com a identificação dos recursos existentes nos softwares de SIG disponíveis no mercado, completam-se os pré-requisitos para a execução da etpa de georeferenciamento da metodologia de pesquisa, descrita no Capítulo 4.

Vale ressaltar a crescente difusão de tecnologias e softwares relacionados direta ou indiretamente ao tema, em particular os sistemas com código aberto, como por exemplo, o Google Earth, permitindo, desta forma, um crescimento significativo desta área do conhecimento; uma vez que um número crescente de pesquisadores e profissionais passam cada vez mais a utilizar tais recursos de forma integrada, tornando a análise (do problema que estiver sendo analisado) muito mais abrangente.

CAPÍTULO 4: METODOLOGIA DE PESQUISA

Este capítulo apresenta um processo para: obtenção de dados não estruturados (textos), a transformação destes dados em informação estruturada e a partir deste novo patamar de informações: realizar a etapa de georeferenciamento, permitindo a partir da combinação destes dois macro-processos, obter conclusões relevantes.

4.1 VISÃO GERAL DA METODOLOGIA:

Estabelecer um processo, com a utilização de Data Mining, Mineração de Textos e Georeferenciamento, para identificar padrões, conceitos e clusters a partir de um banco de dados não estruturados. Na figura 8, pode-se observar o macro-processo da metodologia proposta.

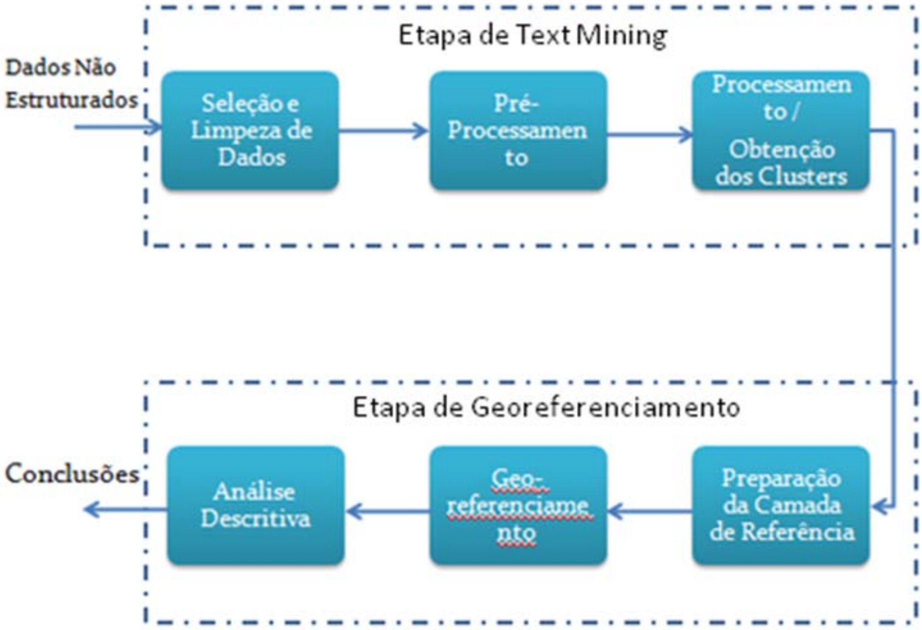


Figura 8: Macro-Processo do Sistema proposto

4.2. A ETAPA DE MINERAÇÃO DE TEXTOS

A etapa de Mineração de Textos pode ser resumida em: "Seleção e Limpeza de Dados", "Pré-Processamento" e "Processamento".

Esquematicamente estas etapas se inter-relacionam como descrito na Figura 9:

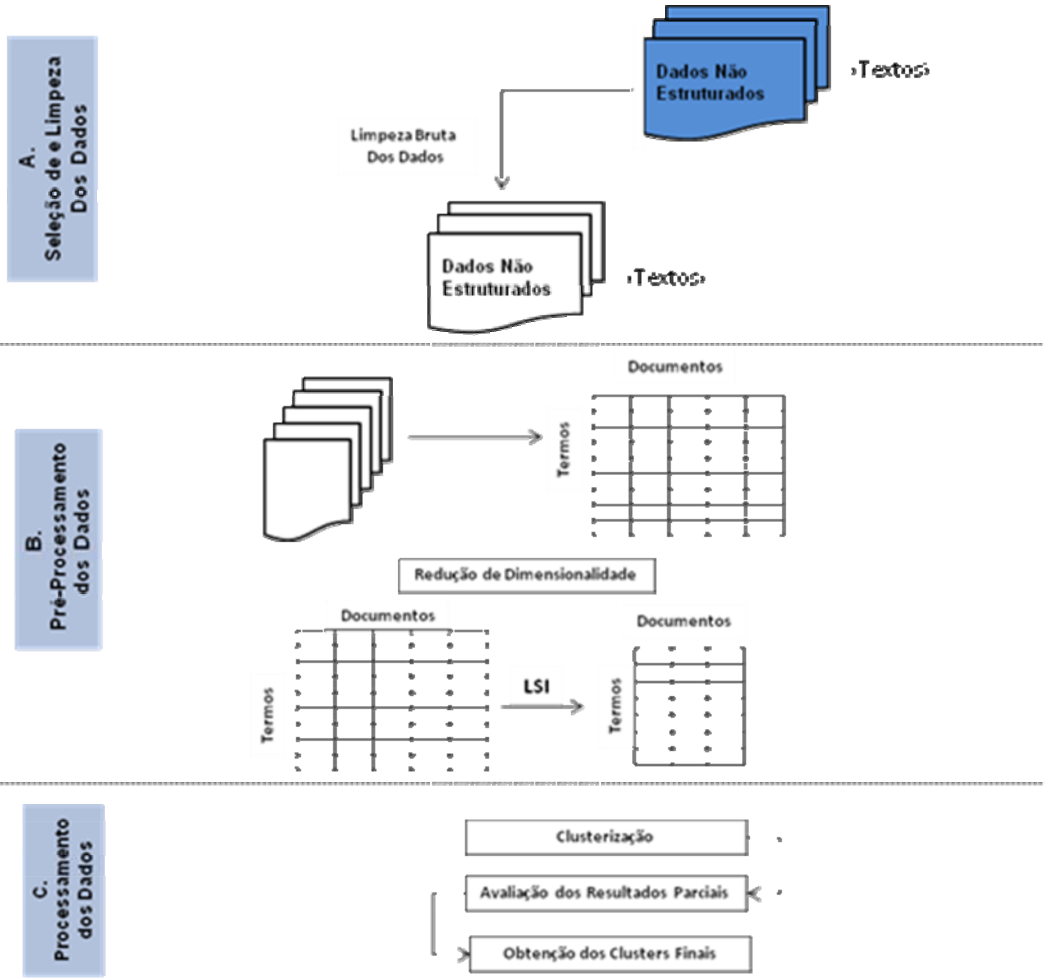


Figura 9: Resumo da Metodologia da Etapa de Mineração de Textos

Resumidamente as etapas mencionadas acima podem ser descritas como:

A) Seleção e Limpeza de Dados: é a extração bruta dos dados onde as primeiras tratativas são efetuadas.

B) Pré-Processamento dos Dados: um conjunto de técnicas e algoritmos que garantem a possibilidade de extração real de conhecimento na etapa posterior, com esforço computacional adequado ao hardware / software disponíveis

C) Processamento dos Dados: execução de procedimentos e algoritmos para a efetiva extração de conhecimento relevante da base de dados textuais.

A seguir o detalhamento das etapas mencionadas:

4.1.1 Etapa A: Extração e Limpeza Inicial da Base de Dados

Nesta etapa, obtém-se como "*output*" uma base preparada para a etapa de pré-processamento, a partir de uma base bruta sem qualquer tratamento, conforme se pode observar esquematicamente na Figura 10.

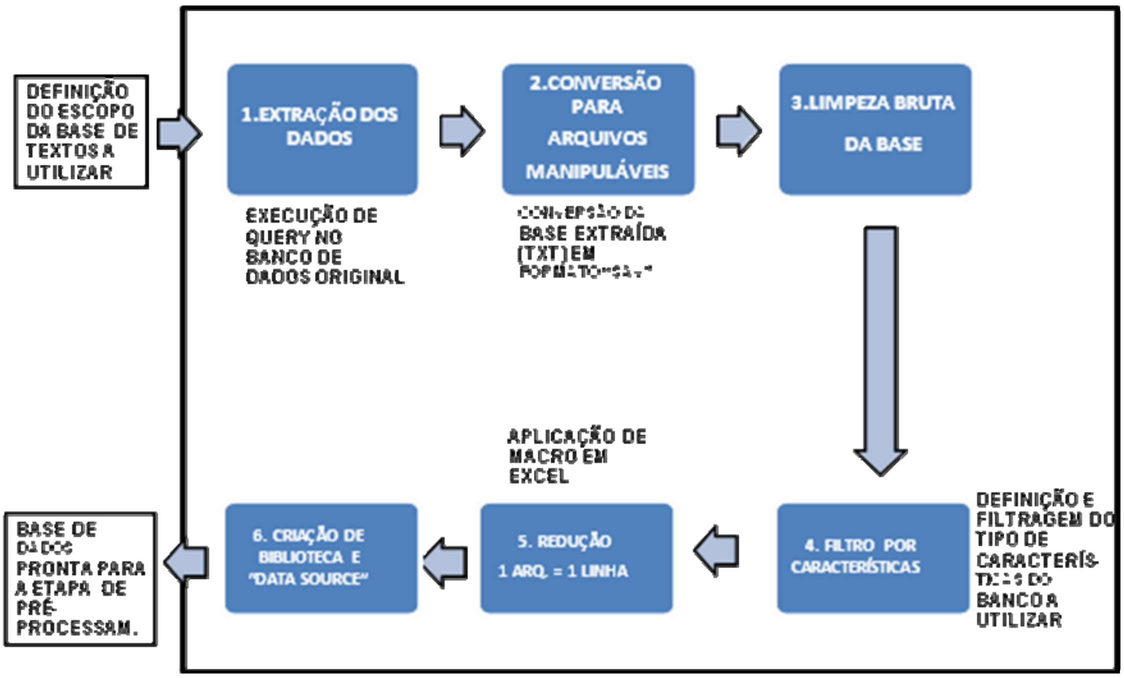


Figura 10: Processo de Extração e Limpeza Bruta dos Dados

1. Extração dos Dados

Nesta etapa do processo, é de fundamental importância conseguir dimensionar a ordem de grandeza do Banco de Dados a analisar, de forma que amostra obtida seja representativa do problema que se queira analisar; e também para que o tempo de processamento de máquina seja adequado. Para o caso de obtenção de dados textuais da Web, existem disponíveis no mercado uma série de aplicativos que fazem esta extração de maneira automatizada: os chamados "*crawlers*". De outra forma, para o caso de bases de dados residente em servidores próprios, utilizam-se *queries* através de recursos convencionais (SQL, por exemplo). Importante na especificação destas *queries* que o Banco de Dados possua não somente a informação textual, como também a informação que servirá como "chave" para o Georeferenciamento.

2. Conversão para Arquivos Manipuláveis

Tanto para o caso de arquivos obtidos na Web, quanto os obtidos em grandes Bancos de Dados, faz-se necessário a conversão da base (de dados) obtida - que nesta etapa do processo, pode ter milhões de linhas - para um arquivo com tamanho reduzido, com ordem de grandeza" de milhares de linhas. Para isto utiliza-se o recurso de amostragem "randômica", ou seja, obtendo uma amostra que represente as características do arquivo original. A partir da análise deste arquivo com as dimensões já reduzidas, algumas métricas básicas podem ser executadas, como por exemplo, a contagem do número de linhas do arquivo gerado e a verificação do número de linhas e campos em branco.

3. Limpeza Bruta

Nesta etapa, caracteres sem representatividade, como por exemplo, %, *, & e @, já podem ser retirados pois não agregarão valor à análise.

4. Filtro por Características

Nesta parte do processo, são filtrados apenas os textos cujas características, por premissas iniciais, virão a ser analisados. Ou seja, caso o Analista esteja trabalhando com uma amostra com muitos tipos de perfis diferentes, este é um bom momento para "filtrar" apenas as características que sofrerão análise. Exemplo: se o Banco de Dados tiver informações textuais de todo o Brasil, mas a análise for realizada somente para um Estado da Federação, este é o momento para retirar da amostra os dados que não serão utilizados na etapa de Processamento Textual, de forma a reduzir o esforço computacional

5. Redução (1 arquivo – 1 linha)

Nesta etapa o objetivo é conseguir transformar a base de dados em um arquivo em que cada linha corresponda a um registro, dado que é muito comum para extrações de bancos de dados textuais o mesmo registro (texto) vir numa sequência de várias linhas consecutivas. Para todos os softwares de Mineração de Textos disponíveis, esta é uma necessidade. Para atingir este objetivo, utiliza-se programação, como a descrita no Apêndice A.

6. Criação de Biblioteca e "*Data Source*"

Para as etapas posteriores de processamento, faz-se necessário a criação de um repositório único, onde as etapas de Mineração de Textos efetivamente serão realizadas, assim de os códigos de programação. Esta também é uma necessidade para a maioria Softwares de Text Mining disponíveis no mercado e permite que o processamento torne-se mais rápido.

4.1.2 Etapa B: Pré-Processamento dos Dados

Este macro-processo pode ser entendido conforme o fluxograma descrito na Fig. 11:

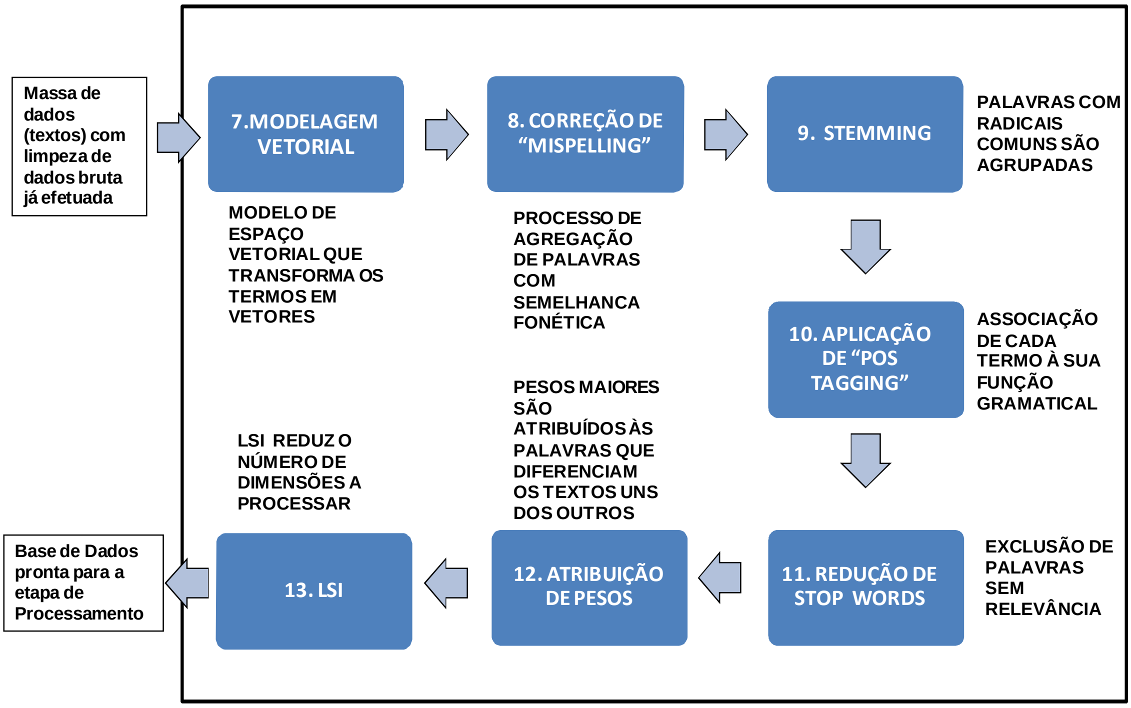


Figura 11: Pré-Processamento dos Dados

Nesta etapa, após a primeira manipulação bruta (realizada na etapa anterior) acontece o Pré-Processamento dos Dados, de forma a organizá-los de uma forma que permita a execução da etapa de Processamento, onde efetivamente as informações relevantes poderão ser extraídas e analisadas.

Ao longo do desenvolvimento da pesquisa vários métodos e algoritmos foram testados nas etapas de 7 a 13 (Pré-Processamento). As etapas a seguir serão listados apenas os métodos onde houve o melhor custo-benefício em termos de (maior) qualidade do pré-processamento x (menor) esforço computacional.

7. Modelagem Vetorial

Nesta etapa faz-se necessário transformar a base de dados em uma matriz padrão do Modelo de Espaço Vetorial (Salton,1975), que representa documentos e consultas como vetores de termos, conforme revisão bibliográfica (item 2.1.2). Termos são ocorrências únicas nos documentos - aos termos são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. O cálculo da similaridade entre os documentos é baseado no ângulo entre os vetores que representam o documento e a consulta.

8. Correção de "*Mispelling*"

De forma a diminuir o esforço computacional e a permitir que palavras que possuam sinônimos sejam contabilizadas como uma única palavra, nesta etapa do Pré-Processamento é utilizado um algoritmo na base de textos exatamente com este objetivo. O *workflow* de atribuição de sinônimos pode ser entendido conforme a Figura 12:

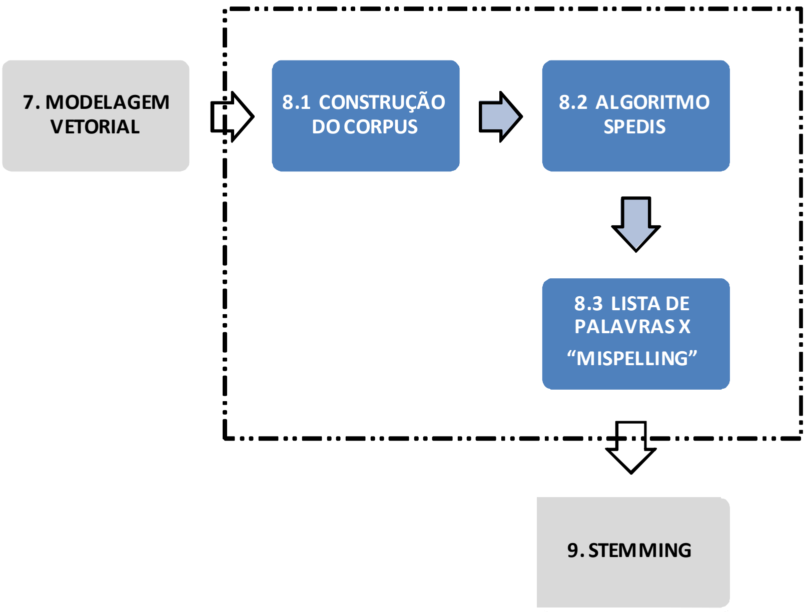


Figura 12: Workflow para criação de Lista de Sinônimos

8.1 Construção do Corpus: o Corpus no idioma Português, foi obtido em www.openoffice.org. Open Office é um software de código aberto para processamento de palavras, planilhas, apresentações, gráficos e bases de dados. Está disponível em vários idiomas e é facilmente processado em computadores de uso pessoal. O download deste Corpus com cerca de trezentas mil palavras, está mais especificamente em:

http://ftp.services.openoffice.org/pub/OpenOffice.org/contrib/dictionaries/pt_BR.zip

8.2 Algoritmo SPEDIS: este algoritmo realiza um cálculo para mensurar a distância fonética entre palavras. A primeira parte do algoritmo SPEDIS pode ser resumidamente descrito da seguinte maneira: supondo que existam duas palavras a comparar. A primeira palavra é a "P1" e a segunda palavra "P2". O algoritmo calcula a distância fonética entre as palavras P1 e P2 ao calcular o "*custo fonético*" entre uma palavra e outra. Este custo fonético é o somatório dos valores que são atribuídos a uma série de operações de comparação entre as palavras analisadas. Ao somar-se todas as operações realizadas nesta operações, tem-se a distância fonética entre as palavras analisadas. A seguir a descrição destas operações e um respectivo exemplo ilustrativo conforme Gershteyn (2005).

Comparando as palavras "SAS" e "SUGI", observando os custos envolvidos na Tabela 5, correspondente ao algoritmo SPEDIS, obtém-se:

- troca da letra U pela letra A (custo =100)

- troca da letra G pela letra S (custo = 100)

- exclusão da letra I do final da palavra (custo = 50)

Portanto custo total é de 250 para esta etapa (primeira parte do algoritmo).

Operação	**Custo**	**Explicação**
match	0	nenhuma mudança
singlet	25	exclusão de um letra dupla
doublet	50	duplicar uma letra
swap	50	inverter a ordem de duas letras consecutivas
truncate	50	excluir a última letra do final da palavra
append	35	incluir uma letra no final da palavra
delete	50	excluir uma letra do meio da palavra
insert	100	inserir uma letra no meio da palavra
replace	100	trocar uma letra no meio da palavra
firstdel	100	excluir a primeira letra da palavra
firstins	200	inserir uma letra no início da palavra
firstrep	200	trocar a priemeira letra da palavra

Tabela 5: Custos das operações do algoritmo SPEDIS

A segunda parte do algoritmo é referente à normalização pela quantidade de caracteres. Desta forma, o cálculo da distância fonética para o exemplo de comparação entre as palavras “SAS” e “SUGI” é o quociente entre o valor do custo total calculado e a quantidade de caracteres da palavra que está sendo comparada, ou seja, no exemplo este valor fica sendo: 250/3 = 83.

A partir do cálculo da distância entre cada conjunto de palavras analisadas, pode-se inferir se duas palavras são foneticamente similares a partir da verificação desta métrica (*distância fonética*). Neste trabalho foi utilizado como referência o valor de “distância fonética” = 15, ou seja, palavras que tenham similaridade até este valor são consideradas como sinônimos.

8.3 Lista de Palavras x “Mispelling”: cada uma das palavras do *dataset* é comparada a cada uma das palavras de um Corpus em português (pode ser obtido grauitamente, por exemplo, a partir do site do www.openoffice.org). A partir desta etapa, o *dataset* já passa a incorporar as relações entre palavras e seus respectivos sinônimos, influenciando desta maneira um novo cálculo de atribuição de pesos.

Se antes co-existiam, por exemplo, as palavras “conhecimento” e “conecimento”, ambas figuravam separadamente com seus respectivos pesos dentro da “Bag of Words”. Agora ambas são representadas por uma única palavra o que vai naturalmente modificar significativamente todo o processo de preparação de dados, uma vez que isto é realizado da mesma forma para todo o conjunto de dados. Se algumas palavras até então não possuíam representatividade isoladamente, podem agora, ao estar agrupadas, ser representativas.

9. Stemming

É o processo de redução das palavras analisadas ao seu respectivo radical, conforme revisão bibliográfica (item 2.1.1). Neste projeto o método de *stemming* utilizado foi o de Porter (1980). Neste processo as palavras do *dataset* são comparadas a um dicionário de palavras já reduzidas ao respectivo radical e após o processamento é possível verificar a associação de cada variação de “*palavra* + *sufixo*” e o respectivo radical associado.

Desta forma, ao reduzir a quantidade de palavras, as palavras variantes de um radical passam a ser encaradas como um único termo.

10. Aplicação de *Part Of Speech Tagging* (*POS Tagging)*

Nesta etapa todas as palavras são categorizadas segundo suas respectivas funções sintáticas permitindo a identificação e consequente exclusão de determinadas funções sintáticas. Conforme o Pré-Processamento vai sendo desenvolvido, novas exclusões podem ser realizadas, inserindo ou excluindo determinadas palavras em função de suas respectivas funções sintáticas. Turney(2002) já mencionava a importância das classes gramaticais das palavras no entendimento de uma série de atributos de uma base de dados textual. Neste trabalho, a utilização do *POS Tagging* tem a função de melhorar a qualidade do agrupamento dos textos na etapa de Processamento.

As palavras foram marcadas segundo as seguintes funções gramaticais: verbos auxiliar, conjunção, artigo, interjeição, preposição, pronome, abreviação, adjetivo, advérbio, numeral, nome, nome próprio, verbo e adjunto adverbial.

Após as diversas interações as seguintes funções gramaticais não foram utilizadas como parte integrante dos *dataset* de Processamento: verbo auxiliar, conjunção, artigo, interjeição, preposição, pronome, abreviação, número e nome próprio.

11. Redução de "*Stop Words*"

Nesta metodologia, a retirada de *stop words* ocorre de duas formas distintas:

- A primeira através do carregamento de listas públicas (por exemplo a que está disponível em www.ranks.nl/stopwords/portuguese.html).

- A segunda forma de redução de Stop Words é realizada a partir da verificação da freqüência das palavras, já com as respectivas associações realizadas nas etapas de *stemming* e de aplicação do algoritmo SPEDIS. Neste trabalho, todas as palavras com freqüência inferior a 3 no *dataset* foram consideradas como Stop Words e portanto não foram utilizadas nas etapas posteriores de processamento.

12. Atribuição de Pesos

Conforme dito anteriormente na etapa inicial do pré-processmento os termos ficam dispostos em uma matriz segundo o Modelo de Espaço Vetorial (Salton, 1975). A partir desta configuração algum método de atribuição de pesos deve ser escolhido, conforme revisão bibliográfica (item 2.1.2). Após diversas interações, o método de atribuição de pesos com melhor performance foi o IDF, sendo:

$$IDF = \log (|D| / DF (t))$$

Onde: |D| é o número total de documentos. Ou seja, quando a IDF de uma palavra ou termo é baixa, se este termo ocorre em muitos documentos. Conseqüentemente IDF tem seu valor máximo quando o termo aparece em apenas um único documento.

13. LSI

Nesta etapa a matriz já com os respectivos pesos atribuídos é reduzida através da técnica LSI (*Latent Semantic Indexing*), conforme Deerwester (1990) e Sebastiani (1999). Este método consiste em organizar uma matriz onde cada coluna corresponde ao histograma da freqüência dos termos obtido em cada documento da coleção para então decompor o espaço expandido pelos vetores coluna em um conjunto ordenado de fatores por um método chamado Decomposição em valores singulares (Singular Value Decomposition-SVD), conforme pesquisa bibliográfica, no item 2.1.3.

A idéia é construir uma matriz de termos por documento na qual seus elementos representam a freqüência de cada termo no documento. A decomposição tem a propriedade de que os últimos fatores têm influência mínima sobre a matriz. Os fatores que menos influenciam podem ser descartados, diminuindo a dimensionalidade. Idealmente, a redução deve ser tal grande o suficiente para permitir que conceitos e detalhes relevantes da base de dados não sejam encobertos. De outra forma, não deve ser grande de forma a inviabilizar o processamento em função do esforço computacional. A técnica utiliza a decomposição em valores singulares (ou SVD-*Singular Value Decomposition*). Essa técnica tem origem em álgebra linear e o resultado é a transformação da matriz de termos original em três outras matrizes. A multiplicação dessas matrizes reconstitui a matriz original.

O passo seguinte do LSI é a redução no tamanho da matriz de valores singulares, ou seja, reduzindo-a de um tamanho (r x r) para um tamanho (k x k), onde k é um número muito menor do que r. Todos os valores além de k serão transformados em zero. Isto faz com que a matriz de valores singulares em passe a ter poucos elementos. Neste trabalho, após diversas interações, o valor de k que permitiu a melhor performance na etapa de Pré-Processamento foi k=100.

4.1.3 Etapa C: Processamento

Nesta etapa é realizada a extração efetiva de conhecimento da base de dados textual, onde diversas técnicas de processamento podem ser aplicadas, conforme Figura 13:

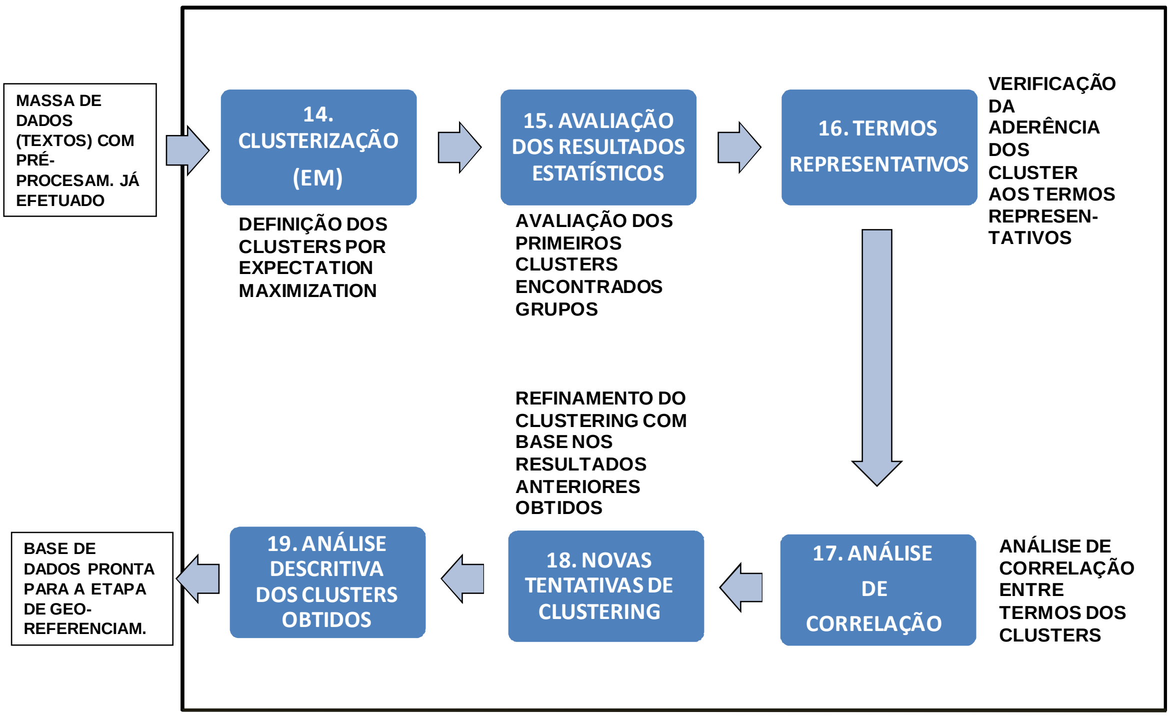


Figura 13: Workflow da etapa de Processamento

14. Clusterização (*Expectation – Maximization*)

Nesta etapa do "*workflow*" é realizado o agrupamento dos documentos por similaridade de características, tendo como "*input*" as variáveis da matriz reduzida pelo processo de LSI (etapa 13). Vários algoritmos (de clustering) podem ser utilizados para este fim, conforme revisão bibliográfica.

Neste trabalho, o método utilizado para agrupamento foi o de *Expectation Maximization* (EM), proposto originalmente por Dempster et al (1977) e descrito em detalhes no Capítulo 2 , pois foi o que permitiu melhores conclusões na etapa de análise (dos clusters finais).

15. Avaliação dos Resultados Estatísticos

Nesta etapa são aplicadas diversas métricas para avaliação da qualidade da separação realizada na etapa de *clusterização*, tais como: verificação da localização dos centróides, quantidade de termos por cluster, freqüência de palavras, freqüência de termos por função sintática, distância dos termos para o centróide, entre outras métricas.

A partir da avaliação estatística nesta etapa é possível avaliar se há necessidade de se continuar a busca de melhores separações (clusters).

16. Termos Representativos

Uma vez obtida a quantidade ideal de clusters, uma importante medida para entendimento dos conceitos associados a cada um destes agrupamentos é a avaliação das palavras representativas destes clusters.

Para identificar os *m* termos de maior representatividade em cada um dos clusters, os *2*m* termos mais freqüentes em cada cluster são utilizados. Para cada um dos *2*m* termos, um cálculo de probabilidade binomial é calculada. A probabilidade de associar um termo a um determinado cluster j é:

$$prob=F(k|N, p)$$

onde:

F é a Função de Distribuição Binomial Cumulativa

k é o número de vezes que o termo aparece no cluster j,

N é o número de documentos no cluster j

p = (sum-k)/(total-N), é a soma do número de vezes que o termo aparece em todos os clusters dividido pelo número total de documentos

Os *m* termos descritivos de cada cluster são aqueles com o maior valor de probabilidade binomial. Neste trabalho foram desconsiderados para os cálculos das probabilidades binomiais os termos com freqüência inferior a 2 por cluster. No capítulo do Estudo de Caso fica bem evidenciada a importância da análise dos termos descritivos para o entendimento das características de cada um dos agrupamentos encontrados.

17. Análise de Correlação

Nesta etapa, outra técnica foi utilizada para análise dos clusters obtidos, a análise de correlação entre os termos, que é uma métrica que quantifica a associação entre termos - também utilizando a distribuição binomial.

Para calcular a força da associação entre um par de termos *a* e *b*, seja:

u o número de documentos contendo o termo *a*
v o número de documentos da coleção
n o número de documentos contend o termo *b*
k o número de documentos contendo tanto o termo *a* quanto o termo *b*
p = k/n a probabilidade do termo *a* ocorrer em conjunto com o termo *b*, assumindo a premissa de que os dois são eventos independentes.

Para determinar a força da associação entre os termos *a* e *b* oriundos de um dataset *r*, primeiramente é computada a distribuição binomial:

Prob(r) = C(n,r) $p^r(1-p)^{(n-r)}$, onde:

C(n,r)= n!/r!(n-r)! é o Coeficiente Binomial

A seguir calcula-se $Prob_k$, sendo:

$$Prob_k = \sum_{r=k}^{r=n} Prob(r)$$

Finalmente, a força da Correlação entre dois termos, a e b é dada por:

Correlação = log20(1/$Prob_k$).

Neste trabalho a intensidade da correlação será evidenciada graficamente, ou seja, a força da correlação entre os termos será demonstrada a partir da grossura da linha que representa a correlação entre dois termos, ou seja, quanto mais grossa a linha mais intensa é a correlação.

18. Novas Tentativas de *Clustering*

As etapas de 14 a 17 são realizadas várias vezes até que os termos representativos (mencionados no item 16) façam sentido, ou seja, denotem conceitos a partir da correlação entre as palavras-chave de cada agrupamento.

19. Análise Descritiva dos Clusters Obtidos

Com base nos termos representativos da etapa 16, nas correlações obtidas na etapa 17 (e tendo como pré-requisito, obtido a separação final, conforme etapa 18) é desejável poder descrever, neste momento, as características globais dos clusters obtidos, ou seja, consolidar as informações e os conceitos de cada cluster, finalizando desta forma a primeiro macro-processo da metodologia proposta (referente a Mineração de Textos).

4.3 A ETAPA DE GEOREFERENCIAMENTO

Nesta etapa, já com os dados textuais processados e com conclusões preliminares a respeito dos clusters obtidos, ocorre a etapa de georeferenciamento, complementando a etapa amterior com novas visões - permitindo desta forma um melhor entendimento do problema estudado. Como premissa para esta etapa faz-se necessário a existência de alguma "*chave*" associada a cada um dos documentos, de forma que estes possam ser georeferencidos. Nesta tese, como será evidenciado no Estudo de Caso, a "chave" é o número do Terminal Fixo de uma Operadora de Telecomunicações - que por este motivo pode ter uma representação no espaço geográfico.

O georeferenciamento é realizado a partir da sobreposição de camadas no espaço geográfico, sendo uma delas a referente à representação dos terminais associados a cada um dos textos processados na etapa de "Text Mining", conforme Figura 14:

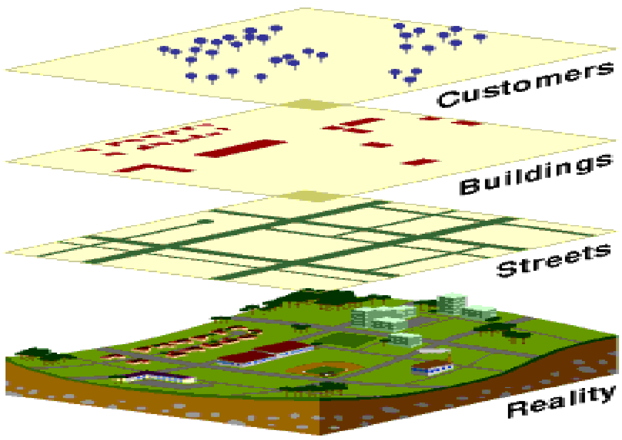


Figura 14: Exemplo de Camadas de Representação em um SIG

4.3.1 Detalhamento do Processo de Georeferenciamento

A seguir, conforme "workflow" representado na Figura 15, são apresentados os processos específicos desta etapa:

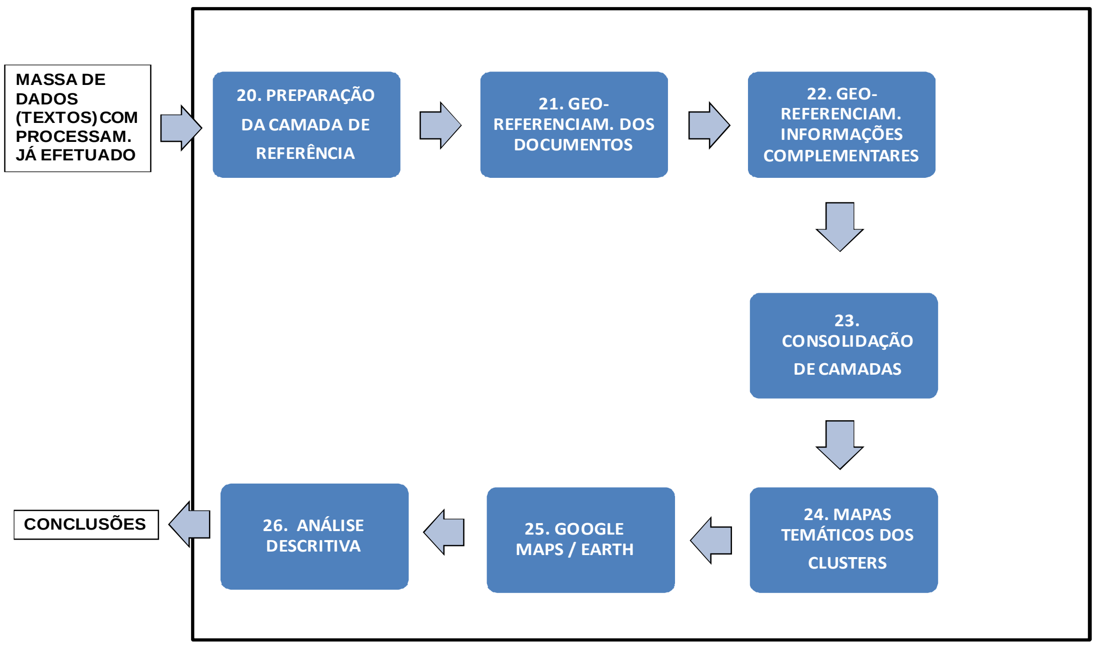


Figura 15: Workflow do Processo de Georeferenciamento

20. Preparação da Camada de Referência

Nesta etapa ocorre o georeferenciamento da camada que será utilizada como referência das demais. Normalemnte algumas camadas (*layers*) são adicionadas sobre esta camada original com o objetivo de qualificação da mesma. Melhor explicando: todas as análises têm como base a camada de referência. Para a criação desta Camada, existem as seguintes possibilidades:

20.1 Utilizando uma mesa digitalizadora: este é o caminho mais demorado para obtenção da Camada de Referência. Caso seja esta a opção, é necessário configurar no SIG o modulo de digitalização e então, selecionar a projeção e o sistema de coordenadas (que não podem ser alteradas depois de começada a digitalização).

20.2 Através da digitalização de uma imagem *raster*[5]: é obter as imagens a partir de um scanner e elaborar os mapas a partir de um cursor, que vai determinar os vetores e polígonos que delimitarão a base de referência

20.3 Importação de uma base já existente em outro SIG: neste tipo de importação faz-se necessário executar algumas rotinas específicas para restabelecer os dados originais e desta forma, algumas informações podem ser perdidas.

20.4 Adquirir uma base cartográfica referente à área de trabalho já perfeitamente tratada: neste caso isso significa que as tabelas vão funcionar com “casamento perfeito” sem perda de informações e sem retrabalho. Contudo normalmente é a opção mais onerosa.

Neste trabalho para a criação da Camada de Referência foram utilizados os métodos 20.2 e 20.4, ou seja, foi adquirida uma base georeferenciada comercial já com tratamento, sobre a qual foram inseridos polígonos e vetores.

---

[5] Imagens raster (ou bitmap, que significa mapa de bits em inglês) são imagens que contém a descrição de cada pixel, em oposição aos gráficos vetoriais. O tratamento de imagens deste tipo requer ferramentas especializadas, geralmente utilizadas em fotografia, pois envolvem cálculos muito complexos, como interpolação, álgebra matricial, etc.

Entendida “a origem” pela qual os dados são obtidos, faz-se necessário definir “a forma” como tais dados serão inseridos no SIG. Para isso, é necessário entender que a maioria dos SIGs funciona como uma base de dados genérica (dados alfanuméricos) que se encontram associados por um identificador comum aos objetos gráficos de um mapa digital. Desta forma um Sistema de Informação Geográfica separa a informação em diferentes camadas temáticas e armazena-as independentemente, permitindo trabalhar com elas de modo rápido e simples, permitindo ao usuário a possibilidade de relacionar a informação existente através da posição e topologia dos objetos, com o fim de gerar uma nova informação. Os modelos mais comuns em SIG são o Modelo Raster ou Matricial e o Modelo Vetorial, descritos a seguir:

- Modelo Matricial: centra-se nas propriedades do espaço, compartimentando-o em células regulares (habitualmente quadradas, mas podendo ser retangulares, triangulares ou hexagonais). Cada célula representa um único valor. Quanto maior for a dimensão de cada célula menor é a precisão ou detalhe na representação do espaço geográfico.

- Modelo Vetorial: o foco das representações centra-se na precisão da localização dos elementos no espaço. Para modelar digitalmente as entidades do mundo real utilizam-se essencialmente três formas espaciais que são o ponto, a linha e o polígono.

Neste trabalho, foi utilizado o Modelo Vetorial, pois a maioria dos softwares de SIG disponíveis utiliza esta forma de representação.

Outra importante parametrização desta etapa (de construção desta primeira camada) e que não pode ser esquecida nas etapas posteriores do workflow de georeferenciamento é a definição do “*Datum*”. Um *Datum* caracteriza-se por uma superfície de referência posicionada em relação à Terra. Um *datum planimétrico* ou horizontal é formalmente estabelecido por cinco parâmetros: dois para definir o elipsóide de referência e três para definir o vetor de translação entre o centro da Terra real e o do elipsóide. Os mapas mais antigos do Brasil adotavam o *datum* planimétrico Córrego Alegre, que utiliza o elipsóide de Hayford. Mais recentemente passou a ser utilizado como referência o Datum SAD-69 que utiliza o elipsóide de referência 1967. A diferença entre os *Datum* são baseadas em modelos matemáticos distintos da forma e dimensões da Terra e do fator adicional da projeção, seja por razões históricas, seja para garantir uma representação gráfica mais proporcionada. Neste trabalho utilizou-se para todos os mapas temáticos gerados, o Datum SAD-69.

21. Georeferenciamento dos Documentos Textuais

Cada documento que passou pelas etapas do workflow de Mineração de Textos precisa necessariamente ter alguma "*chave*" que permita georeferenciá-los individualmente. Para a obtenção das coordenadas *x* e *y* neste trabalho foram utilizadas duas formas distintas para isto:

- a primeira, mais simples, a partir do "acoplamento" desta base de documentos a uma base de referência;

- a segunda, a partir da informação do CEP (Código de Endereçamento Postal) e de programação em KML (para a obtenção das coordenadas), para o georeferenciamento de camadas complementares, com visões obtidas a partir de fotos de satélites (conforme descrito no item 3.5).

A seguir as duas propostas mencionadas:

21.1 Georeferenciamento dos documentos baseado no acoplamento de tabelas: neste caso deve-se associar a tabela referente aos documentos com a tabela que contém os dados da camada de referência. Uma das capacidades do software SIG é a de dispor de ferramentas que permitem gerir, tratar e manipular bases de dados.

Especificamente para inserir esta nova informação, ou seja, o georeferenciamento de cada um dos textos, adotou-se a solução de fazer com que estes dados passem a fazer parte da camada de referência e não como uma nova camada sobreposta a outras. Isto funciona como um procedimento de manutenção da tabela de dados existente, ou seja, uma nova coluna nesta tabela (da camada de referência) passa a existir. Feita a atualização da tabela, esta já pode ser visualizada no mapa com a informação dos documentos georeferenciados.

21.2 Georeferenciamento baseado nos CEPs: no caso de não existir uma “chave”comum entre a tabela de dados da camada de referência e a os documentos oriundos da etapa de Text Mining, faz-se necessário georeferenciar (tais documentos) através da criação de uma nova camada. A seguir será demonstrada uma alternativa para estabelecer tais coordenadas a partir do Código de Endereçamento Postal, conforme fluxo da Figura 16.

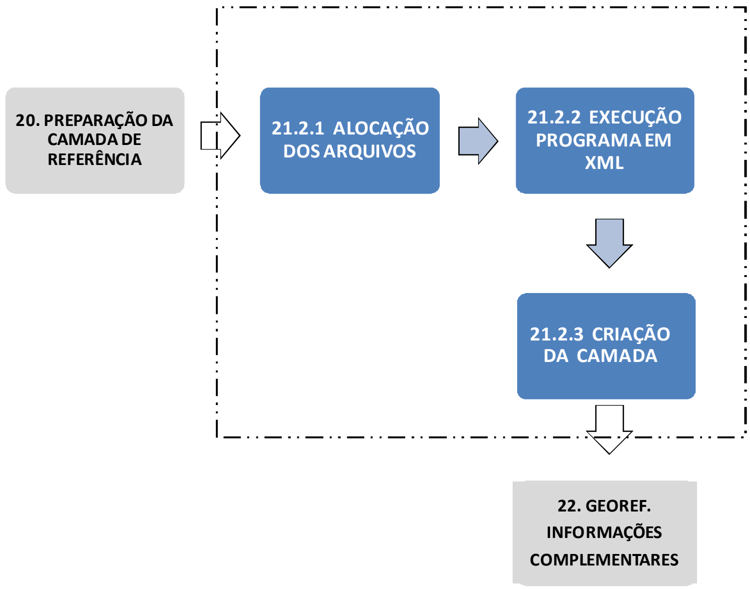


Figura 16: Georeferenciamento baseado no CEP

21.2.1 Alocação dos arquivos: nesta etapa é necessário dispor os documentos e o programa em XML em um mesmo diretório em dois arquivos separados. O arquivo que contém as referências dos documentos deve estar em formato txt e a informação do CEP deve vir conforme o exemplo a seguir:

Exemplo de Ordenação de Ponto a Georeferenciar x CEP:

```
<?xml version="1.0"?>
<pontos>
<ponto cep="60325-003, CE, BRA"/>
<ponto cep="60055-210, CE, BRA"/>
<ponto cep="60150-160, CE, BRA"/>
<ponto cep="60115-282, CE, BRA"/>
</pontos>
```

21.2.2 Execução do Programa em XML: nesta etapa o programa em XML que utiliza a função *geocoder.getLocations* (ver Apêndice G), é executado, o que irá gerar as coordenadas *x* e *y*, que é o objetivo desta etapa. Após a execução deste programa observa-se no exemplo a seguir o "output" da programção em XML.

Exemplo de *Output* (informações em negrito são as coordenadas geradas pelo programa):

<Placemark><styleUrl>#idstylecep</styleUrl><name>60325-003,CE,
BRA</name><Point><coordinates>**-38.5605081,-3.7249144**</coordinates></Point></Placemark>
<Placemark><styleUrl>#idstylecep</styleUrl><name>60055-210,CE,
BRA</name><Point><coordinates>**-38.5230905,-3.7462091**</coordinates></Point></Placemark>
<Placemark><styleUrl>#idstylecep</styleUrl><name>60150-160,CE,
BRA</name><Point><coordinates>**-38.5132387,-3.7314852**</coordinates></Point></Placemark>
<Placemark><styleUrl>#idstylecep</styleUrl><name>60115-282,CE,
BRA</name><Point><coordinates>**-38.511926,-3.7472959**</coordinates></Point></Placemark>

Este output da forma como está pode ser salvo como um arquivo KML para georeferenciamento no Google Earth. Porém nesta etapa apenas as informações das coordenadas são utilizadas (em negrito no exemplo acima).

21.2.3 Criação da Camada: nesta etapa as coordenadas *x* e *y* obtidas são alocadas em uma planilha e então inseridas no SIG. Desta forma os pontos são criados na representação do espaço geográfico, passando a representar uma camada individualizada que pode agora ser sobreposta às demais.

22. Georeferenciamento de Informações Complementares

Nesta etapa são criadas camadas a partir de informações sócio-demográficas que ajudarão ao usuário tecer conclusões quando os clusters estiverem definidos em sua representação geográfica. Informações como por exemplo, classificação social, quantidade de domicílios, população, entre outras variáveis, podem ser inseridas neste contexto. Pode-se optar por adquirir tais bases comercialmente, já georeferenciadas, ou construídas a partir de dados que contenham alguma chave que possa gerar as coordenadas *x* e *y* de maneira análoga ao descrito no item 21.1.

23. Consolidação das Camadas

Nesta etapa ocorre a manipulação das camadas no SIG de forma a ordená-las de uma forma que seja possível a posterior construção e visualização do mapa temático dos clusters.

Desta forma cada uma das camadas pode ser gerenciada de forma a modificar-se a ordenação das mesmas, ou seja, qual camada deve aparecer sobre outra, quais *labels* deverão ou não aparecer, qual a escala das fontes utilizadas, qual o *zoom* do mapa, etc.

24.Mapas Temáticos dos Clusters

Nesta etapa ocorre a representação dos Clusters sob a forma de um "geo-campos temáticos" ou simplesmente "mapa temático".

Segundo Monteiro & Camara (2007):

> *"Um geo-campo representa a distribuição espacial de uma variável que possui valores em todos os pontos pertencentes a uma região geográfica, num dado tempo t".*

Os geo-campos podem ser especializados em: "Temático", "Numérico" e de "Dado Sensor Remoto". Neste trabalho é utilizado para a representação dos Clusters o Geo-Campo Temático, onde: dada uma região geográfica R, associa a cada ponto do espaço um tema de um mapa. Na Figura 17 um exemplo de representação temática de tipos diferentes de solo.

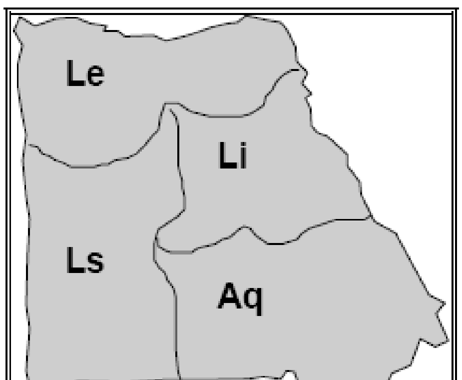


Figura 17: Exemplo de representação temática

Nesta Tese, a representação temática é realizada a partir da identificação (da sobreposição) da representação geográfica dos documentos sobre a camada de referência.

Percorridas todas as etapas do item 24, obtém-se o mapa temático em uma das camadas sobreposta às demais existentes, permitindo já a partir deste momento as primeiras análises e conclusões a respeito do tema (neste caso os clusters de documentos).

25. Google Maps / Google Earth

A partir do georeferenciamento e interligação dos vértices de cada um dos polígonos da camada que se queira construir tendo como "pano de fundo" o Google Maps / Google Earth, é possível visualizar os dados das camadas construídas nas etapas anteriores sob uma nova ótica, a partir do carregamento de um programa KML diretamente na *homepage* da Google (ver Apêndice G).

4.4 RESUMO DO CAPÍTULO

Neste capítulo foi possível evidenciar a principal contribuição desta Tese, ou seja, o detalhamento de um processo de Mineração de Textos combinada ao Georeferenciamento, com as respectivas integrações de ferramentas e algoritmos - tudo isto suportado pela revisão bibliográfica pertinente.

Ao utilizar a ordem processual descrita no capítulo, é possível, sem dúvida alguma, aplicar tais conhecimentos em uma situação de Estudo de Caso, conforme será verificado no Capítulo 5.

Porém, mais do que permitir a execução deste Estudo de Caso, a relevância de relacionar cada uma das etapas realizadas é: permitir que futuras pesquisas nesta área possam ser realizadas utilizando o conhecimento adquirido a partir deste trabalho.

CAPÍTULO 5: O ESTUDO DE CASO

Nesta etapa encontra-se a aplicação da metodologia de pesquisa em um Estudo de Caso. A seguir o detalhamento das atividades realizadas e do problema estudado.

5.1 O BANCO DE DADOS TEXTUAL

Obteve-se uma base de dados textuais de registros de clientes de Telefonia Fixa - de uma empresa prestadora de serviços de Telecomunicações – de Consumidores que entraram em contato com a área de Atendimento (*Call Center*). De uma base de cerca de quinze milhões de registros referente a nove meses de atendimento, foram selecionados registros ao longo dos meses de janeiro a setembro de 2008. Deste Banco de Dados original, foram selecionados aleatoriamente (com auxílio de software) cerca de 65.000 linhas, facilitando desta forma a conversão entre arquivos entre diferentes softwares.

A manipulação inicial da base de 65.000 linhas foi feita em Excel - para uma primeira limpeza de dados, ainda grosseira, de caracteres indesejáveis e pontuação. Posteriormente, com a execução das demais etapas (de Pré-Processamento), este valor foi reduzido a um arquivo de cerca de 2500 linhas, cada uma destas (linhas) correspondendo a um único registro de cliente (dado que os registros ocupavam muitas linhas consecutivas). Outro motivo que fez este arquivo original ser reduzido nesta proporção foi o fato de terem sido “filtrados” apenas registros de um mesmo tipo de produto específico desta Operadora: um produto de telefonia com consumo mínimo de 50 reais por mês por cliente.

Ainda como limitador da análise, de forma a reduzir o esforço computacional (realizado em um desktop) foram selecionados apenas terminais de clientes com DDD 21, ou seja, pertencentes ao Rio de Janeiro (Capital e Grande Rio, incluindo a Baixada Fluminense e as regiões de Niterói e São Gonçalo).

Na Figura 18 observa-se uma amostra do Banco de Dados mencionado, omitindo-se as colunas referentes ao número do terminal e da conta dos clientes, em função da confidencialidade dos dados pessoais destes (condição exigida pela Operadora que cedeu o *dataset*). Nota-se, como descrito anteriormente, o texto ainda com muitos caracteres sem importância para o processo de Mineração de Textos e a formatação de cada registro ainda em várias linhas consecutivas.

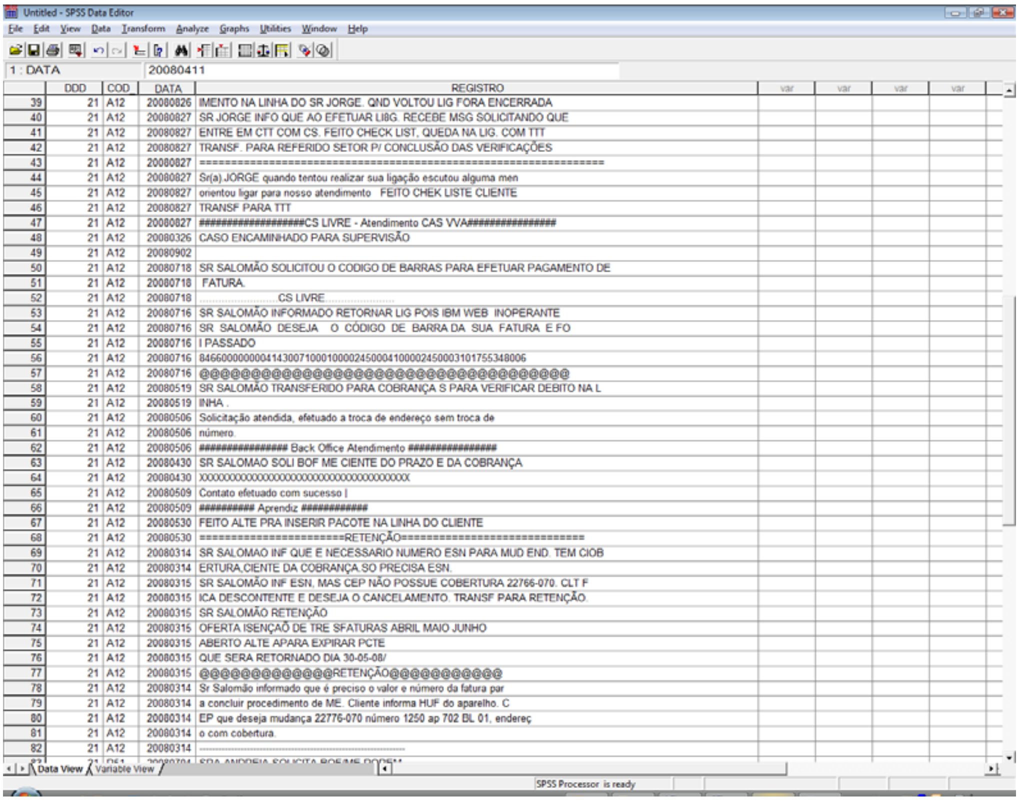


Figura 18: Amostra do Dataset Original

5.2 O PRÉ –PROCESSAMENTO DA BASE

Para a criação da Biblioteca, com o repositório dos arquivos textuais e para a realização das etapas de pré-processamento e processamento (conforme metodologia, item 4.1.1, sub-item 6) foi executada programação (ver Apêndice B). Na Figura 19 observa-se o log dos arquivos gerados.

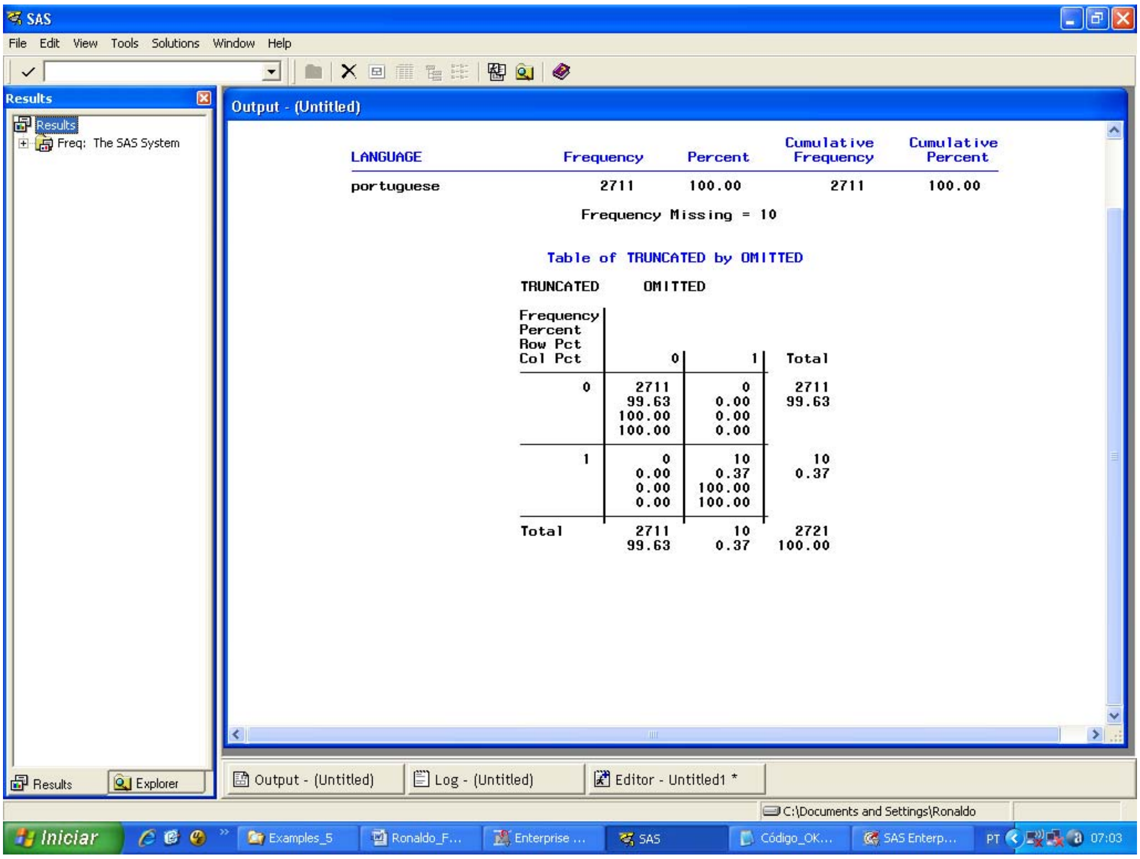


Figura 19: Log de Geração de Arquivo Textual para etapa de Pré-Processamento

Observação: Os 10 registros que aparecem como “truncated” são os registros de clientes que estavam em branco, ou seja, havia a marcação do número de um terminal e respectivas identificações de um arquivo de registro de cliente, mas com informação textual em branco.

Na Figura 20 observa-se um “*Print Screen*” do arquivo no formato já preparado para a etapa de Pré-Processamento:

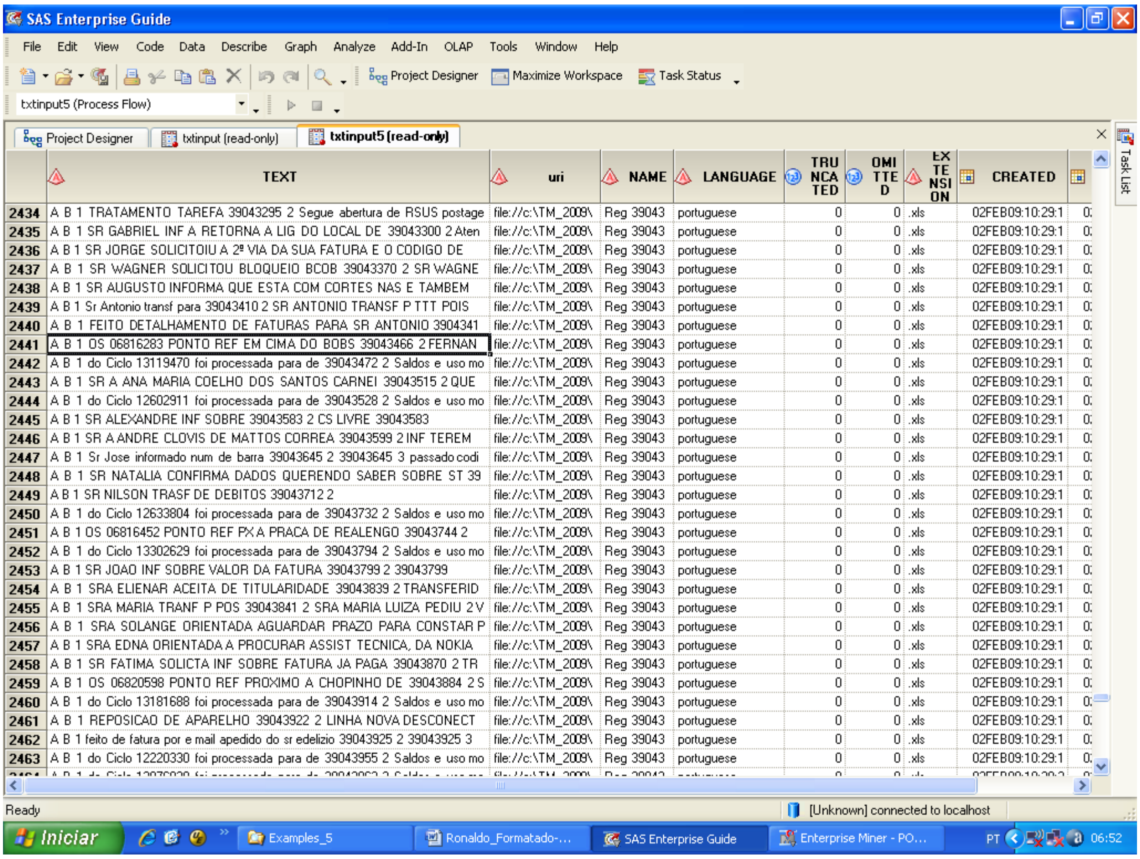


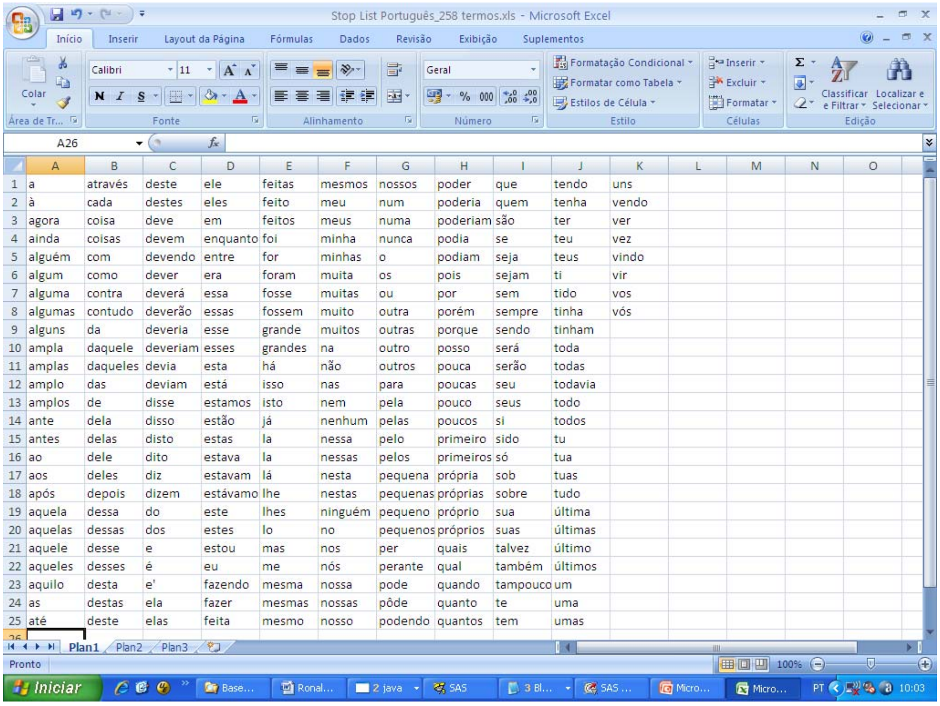
Figura 20: Amostra do Banco de Dados Textual

Nos itens a seguir a descrição das tarefas realizada no Pré-Processamento propriamente dito:

A Geração de Stop Words

A construção da referida Stop List foi realizada a partir de um dataset representativo da base original (2517 documentos, como dito anteriormente), mas foi sendo aperfeiçoada a medida que os refinamentos dos algoritmos de *parsing* e *stemming* foram aplicados. Ou seja, como em um processo de Mineração de Dados típico, os aperfeiçoamentos, inclusive da lista de Stop Words, foi feita de maneira interativa, ou seja, conforme as melhorias nos proceso de Parsing, Stemming e de associação de sinônimos foram acontecendo, novas listas de Stop Words foram atualizadas. Até a formação dos Clusters Finais, que serão apresentados na sequência desta Tese, cerca de 30 atualizações da Lista de Stop Words foram realizadas. Além disso, por premissa adotada, os termos que apareceram com freqüência igual ou menor que 2, foram eliminados de maneira primária ou seja, entraram na Stop List. Adicionalmente, artigos, conjunções, abreviações, números e nomes próprios, ou seja, palavras que não agregam valor à discriminação ou diferenciação dos textos entre si, foram eliminadas do *dataset*.

Além da exclusão das palavras sob os critérios descritos acima, foram adicionadas à Stop List um conjunto de 258 termos adicionais, em português, conforme Metodologia de Pesquisa, item 4.1.2. Na Figura 21 observa-se a listagem de 258 palavras mencionada.



	A	B	C	D	E	F	G	H	I	J	K
1	a	através	deste	ele	feitas	mesmos	nossos	poder	que	tendo	uns
2	à	cada	destes	eles	feito	meu	num	poderia	quem	tenha	vendo
3	agora	coisa	deve	em	feitos	meus	numa	poderiam	são	ter	ver
4	ainda	coisas	devem	enquanto	foi	minha	nunca	podia	se	teu	vez
5	alguém	com	devendo	entre	for	minhas	o	podiam	seja	teus	vindo
6	algum	como	dever	era	foram	muita	os	pois	sejam	ti	vir
7	alguma	contra	deverá	essa	fosse	muitas	ou	por	sem	tido	vos
8	algumas	contudo	deverão	essas	fossem	muito	outra	porém	sempre	tinha	vós
9	alguns	da	deveria	esse	grande	muitos	outras	porque	sendo	tinham	
10	ampla	daquele	deveriam	esses	grandes	na	outro	posso	será	toda	
11	amplas	daqueles	devia	esta	há	não	outros	pouca	serão	todas	
12	amplo	das	deviam	está	isso	nas	para	poucas	seu	todavia	
13	amplos	de	disse	estamos	isto	nem	pela	pouco	seus	todo	
14	ante	dela	disso	estão	já	nenhum	pelas	poucos	si	todos	
15	antes	delas	disto	estas	la	nessa	pelo	primeiro	sido	tu	
16	ao	dele	dito	estava	la	nessas	pelos	primeiros	só	tua	
17	aos	deles	diz	estavam	lá	nesta	pequena	própria	sob	tuas	
18	após	depois	dizem	estávamo	lhe	nestas	pequenas	próprias	sobre	tudo	
19	aquela	dessa	do	este	lhes	ninguém	pequeno	próprio	sua	última	
20	aquelas	dessas	dos	estes	lo	no	pequenos	próprios	suas	últimas	
21	aquele	desse	e	estou	mas	nos	per	quais	talvez	último	
22	aqueles	desses	é	eu	me	nós	perante	qual	também	últimos	
23	aquilo	desta	e'	fazendo	mesma	nossa	pode	quando	tampouco	um	
24	as	destas	ela	fazer	mesmas	nossas	pôde	quanto	te	uma	
25	até	deste	elas	feita	mesmo	nosso	podendo	quantos	tem	umas	

Figura 21: Listagem de Palavras adicionadas à Stop Words List

A partir desta Listagem de Stop Words, mais completa, pode-se observar através de uma verificação simples da freqüência das palavras, que muitos termos que apareciam com freqüência baixa poderiam ser agrupados ou substituídos por seus respectivos sinônimos. Por este motivo, uma nova etapa no Pré-Processamento foi realizada na sequência: a criação de uma lista de palavras com semelhança fonética, que foi utilizada como comparação a cada uma das palavras do *dataset* trabalhado, conforme Metodologia de Pesquisa (item 4.1.2, sub-item 8). Para a criação da lista de palavras com semelhança fonética, utilizou-se o algoritmo SPEDIS, através de programação (ver Apêndice E).

Com esta lista de palavras foneticamente equivalentes já aplicada, um novo *parsing* foi realizado e a lista de *stop words* tornou-se mais concisa – dado que vários termos até então considerados com freqüência igual ou inferior a 2, passaram a ter um freqüência maior por estarem agrupados com seus respectivos sinônimos - e por este motivo deixaram de constar na S*top List*.

Thesaurus

A seguir a descrição dos procedimentos e programação para a criação de um Thesaurus e o respectivo Dicionário de Sinônimos. O Corpus no idioma Português, foi obtido em www.openoffice.org.[6] O download deste Corpus com cerca de trezentas mil palavras, foi obtido em:

http://ftp.services.openoffice.org/pub/OpenOffice.org/contrib/dictionaries/pt_BR.zip

Na Figura 22 observa-se uma amostra do Corpus mencionado.

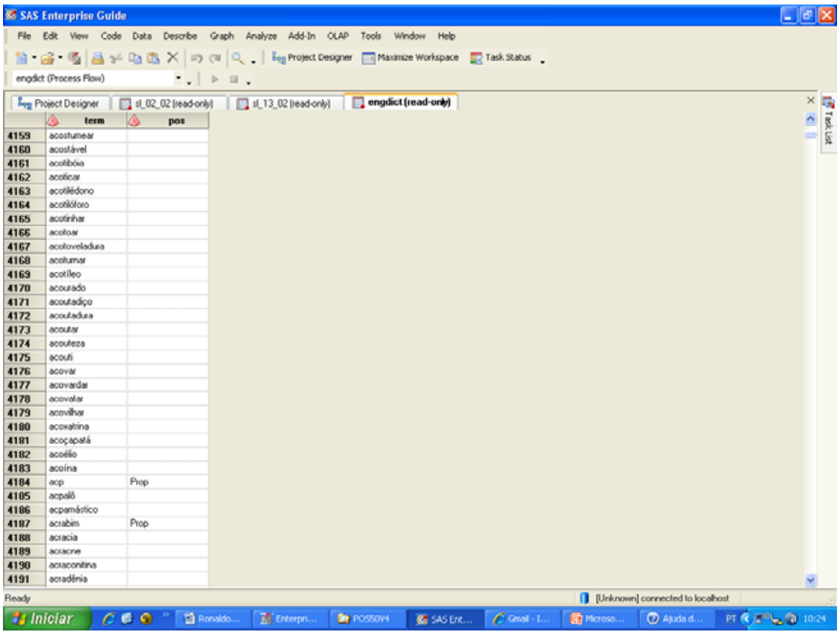


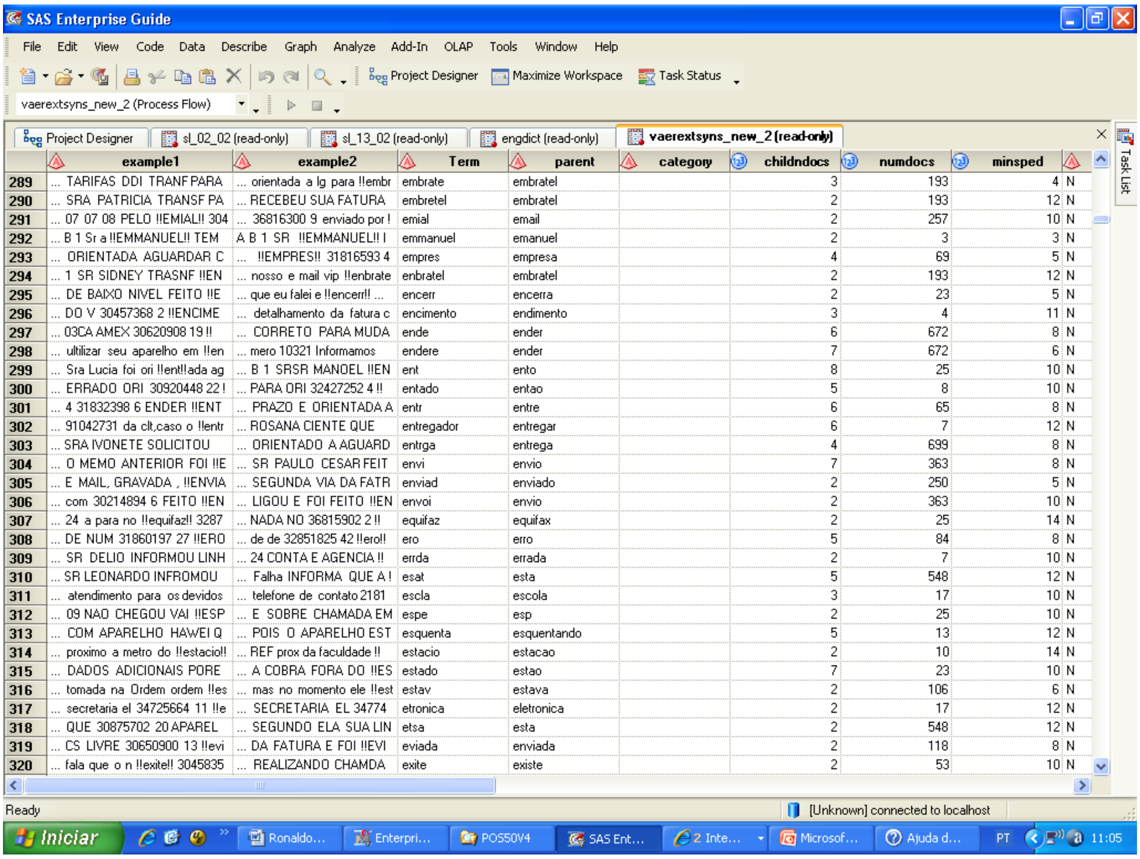
Figura 22: Amostra do *Corpus*

Para que os termos do *dataset* original (2517 documentos) pudessem ter sua correspondência em relação ao dicionário de Português baixado, foi utilizada

[6] OpenOffice.org é um software de código aberto para procesamento de palavras, planilhas, apresentações, gráficos e bases de dados. Está disponível em vários idiomas e roda facilmente em computadores de uso pessoal

programação no SAS Base para a criação de uma correlação, baseada na fonética de cada uma das palavras. Para isto foi executada programação, conforme Apêndice E.

Na Figura 23, observa-se a lista de sinônimos já criada e aplicada ao dataset:



	example1	example2	Term	parent	category	childndocs	numdocs	minsped	
289	... TARIFAS DDI TRANF PARA	... orientada a lg para !!embr	embrate	embratel		3	193	4	N
290	... SRA PATRICIA TRANSF PA	... RECEBEU SUA FATURA	embretel	embratel		2	193	12	N
291	... 07 07 08 PELO !!EMIAL!! 304	... 36816300 9 enviado por !	emial	email		2	257	10	N
292	... B 1 Sr a !!EMMANUEL!! TEM	A B 1 SR !!EMMANUEL!! I	emmanuel	emanuel		2	3	3	N
293	... ORIENTADA AGUARDAR C	... !!EMPRES!! 31816593 4	empres	empresa		4	69	5	N
294	... 1 SR SIDNEY TRASNF !!EN	... nosso e mail vip !!enbrate	enbratel	embratel		2	193	12	N
295	... DE BAIXO NIVEL FEITO !!E	... que eu falei e !!encerr!! ...	encerr	encerra		2	23	5	N
296	... DO V 30457368 2 !!ENCIME	... detalhamento da fatura c	encimento	endimento		3	4	11	N
297	... 03CA AMEX 30620908 19 !!	... CORRETO PARA MUDA	ende	ender		6	672	8	N
298	... ultilizar seu aparelho em !!en	... mero 10321 Informamos	endere	ender		7	672	6	N
299	... Sra Lucia foi ori !!ent!!ada ag	... B 1 SRSR MANOEL !!EN	ent	ento		8	25	10	N
300	... ERRADO ORI 30920448 22 !	... PARA ORI 32427252 4 !!	entado	entao		5	8	10	N
301	... 4 31832398 6 ENDER !!ENT	... PRAZO E ORIENTADA A	entr	entre		6	65	8	N
302	... 91042731 da clt,caso o !!entr	... ROSANA CIENTE QUE	entregador	entregar		6	7	12	N
303	... SRA IVONETE SOLICITOU	... ORIENTADO A AGUARD	entrga	entrega		4	699	8	N
304	... O MEMO ANTERIOR FOI !!E	... SR PAULO CESAR FEIT	envi	envio		7	363	8	N
305	... E MAIL, GRAVADA , !!ENVIA	... SEGUNDA VIA DA FATR	enviad	enviado		2	250	5	N
306	... com 30214894 6 FEITO !!EN	... LIGOU E FOI FEITO !!EN	envoi	envio		2	363	10	N
307	... 24 a para no !!equifaz!! 3287	... NADA NO 36815902 2 !!	equifaz	equifax		2	25	14	N
308	... DE NUM 31860197 27 !!ERO	... de de 32851825 42 !!ero!!	ero	erro		5	84	8	N
309	... SR DELIO INFORMOU LINH	... 24 CONTA E AGENCIA !!	errda	errada		2	7	10	N
310	... SR LEONARDO INFROMOU	... Falha INFORMA QUE A !	esat	esta		5	548	12	N
311	... atendimento para os devidos	... telefone de contato 2181	escla	escola		3	17	10	N
312	... 09 NAO CHEGOU VAI !!ESP	... E SOBRE CHAMADA EM	espe	esp		2	25	10	N
313	... COM APARELHO HAWEI Q	... POIS O APARELHO EST	esquenta	esquentando		5	13	12	N
314	... proximo a metro do !!estacio!!	... REF prox da faculdade !!	estacio	estacao		2	10	14	N
315	... DADOS ADICIONAIS PORE	... A COBRA FORA DO !!ES	estado	estao		7	23	10	N
316	... tomada na Ordem ordem !!es	... mas no momento ele !!est	estav	estava		2	106	6	N
317	... secretaria el 34725664 11 !!e	... SECRETARIA EL 34774	etronica	eletronica		2	17	12	N
318	... QUE 30875702 20 APAREL	... SEGUNDO ELA SUA LIN	etsa	esta		2	548	12	N
319	... CS LIVRE 30650900 13 !!evi	... DA FATURA E FOI !!EVI	eviada	enviada		2	118	8	N
320	... fala que o n !!exite!! 3045835	... REALIZANDO CHAMDA	exite	existe		2	53	10	N

Figura 23: Lista de Sinônimos aplicada ao Dataset Original

Na Figura acima, a coluna "*Term*" apresenta o termo analisado e a coluna "Parent" apresenta o termo existente no dicionário previamente carregado. Como foi dito, esta relação passa a fazer parte do fluxo de processamento e permitindo ganho expressivo de produtividade no *parsing*, quando este foi realizado novamente, o que posteriormente melhorou o processo de Clusterização.

O Processo de *Stemming*

Tendo as etapas anteriores sido realizadas, com aprimoramento do Parsing, cabe ressaltar e exemplificar também como etapa fundamental, o processo de Stemming. Na Figura 24, pode-se verificar a palavra "cancelar" e todas as suas variantes (cancelado, cancelou, cancelados, cancela, cancelada) que passaram a ser contabilizados como sendo um único termo, para efeito de *parsing*.

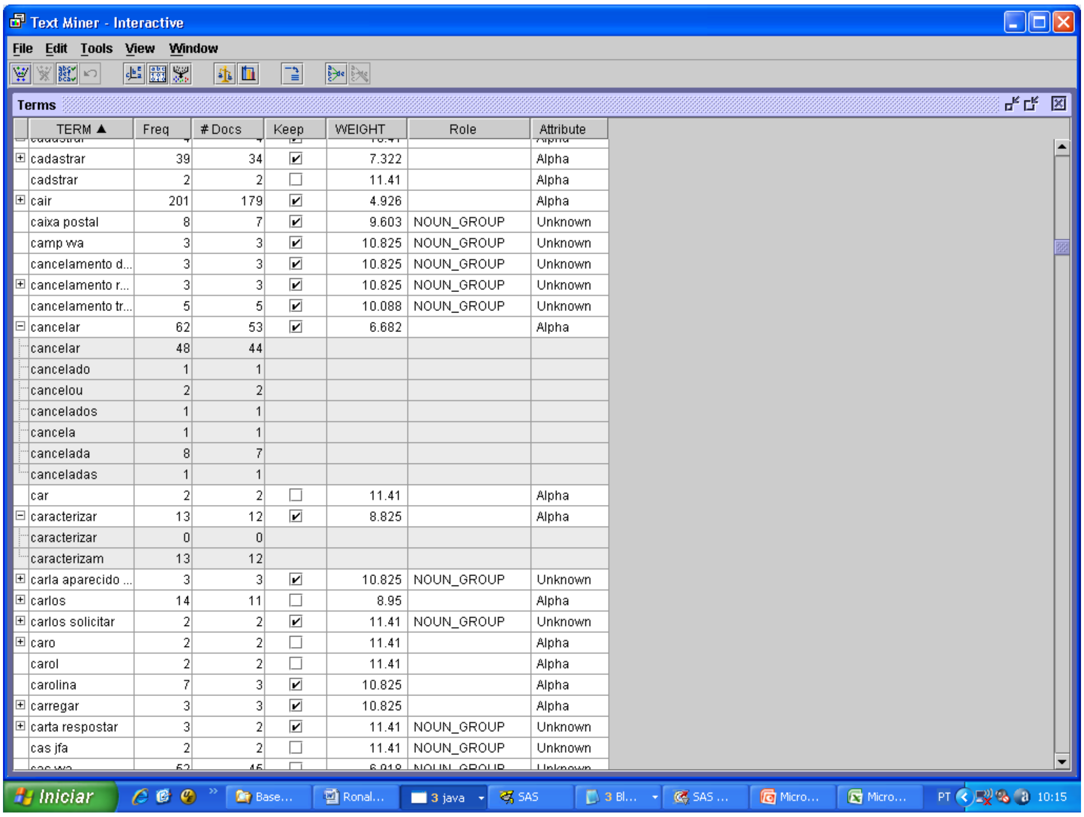


Figura 24: Exemplo de Algoritmo de Stemming aplicado ao Dataset

A Atribuição de Pesos aos Termos

Uma etapa importante para uma adequada finalização da etapa de Pré-Processamento é o de Atribuição de Pesos aos Termos, permitindo uma adequada representação de cada documento do *dataset*.

Na Figura 25, observa-se a *plotagem* dos Termos em função de: "Frequência no Dataset" x "Quantidade de Documentos em que os mesmos aparecem". É a partir da lógica deste binômio de quantidade de termos no *dataset* e nos documentos que a etapa seguinte (de atribuição de pesos) irá usar como input para o cálculo de atribuição de pesos.

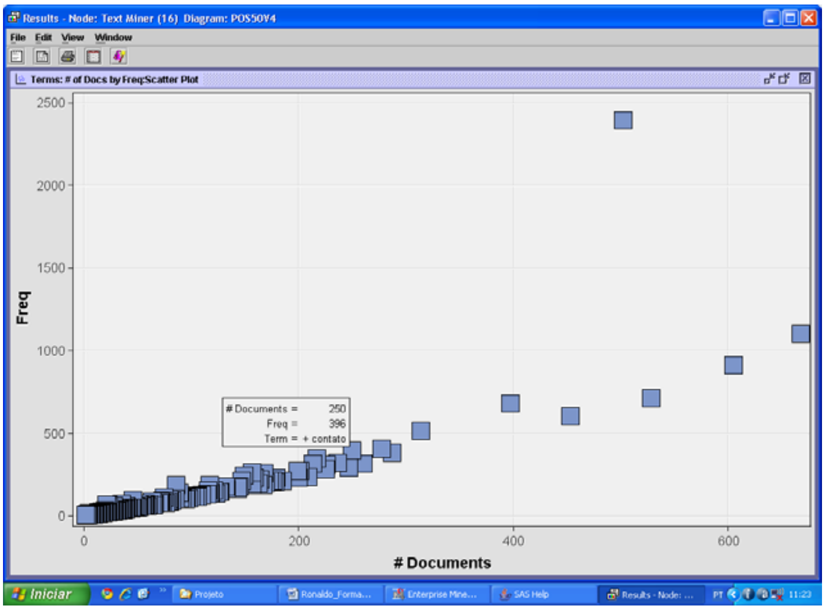
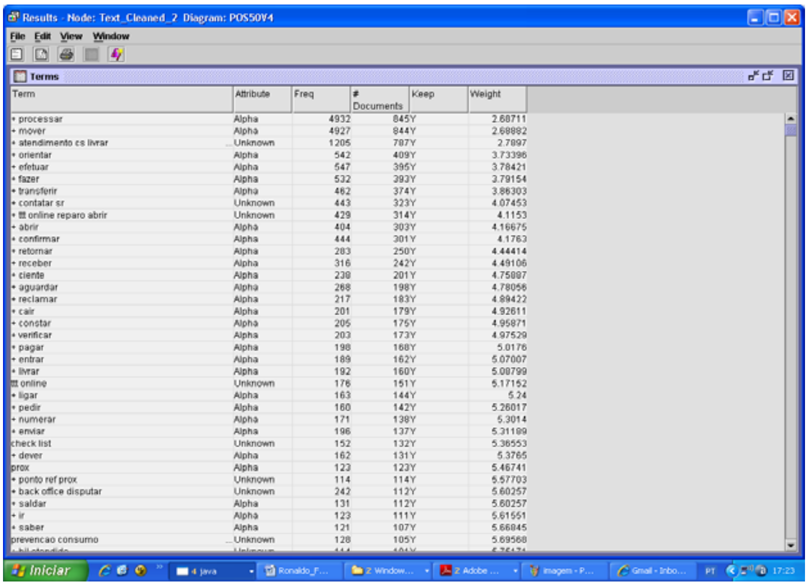


Figura 25: Plotagem das Frequências x Quantidade de Documentos

Como já citado anteriormente, sendo o processo de Data Mining interativo, vários métodos de atribuição de pesos foram testados e o que permitiu uma melhor separação dos clusters foi o método IDF. A seguir, na Figura 26 a visualização de uma amostra de atribuição de pesos aos respectivos termos com a utilização do algoritmo IDF.


Results - Node: Text_Cleaned_2 Diagram: POS50V4

Terms

Term	Attribute	Freq	# Documents	Keep	Weight
+ processar	Alpha	4932	845	Y	2.68711
+ mover	Alpha	4927	844	Y	2.68882
+ atendimento cs livrar	Unknown	1205	787	Y	2.7897
+ orientar	Alpha	542	409	Y	3.73396
+ efetuar	Alpha	547	395	Y	3.78421
+ fazer	Alpha	532	393	Y	3.79154
+ transferir	Alpha	462	374	Y	3.86303
+ contatar sr	Unknown	443	323	Y	4.07453
+ ttt online reparo abrir	Unknown	429	314	Y	4.1153
+ abrir	Alpha	404	303	Y	4.16675
+ confirmar	Alpha	444	301	Y	4.1763
+ retornar	Alpha	283	250	Y	4.44414
+ receber	Alpha	316	242	Y	4.49106
+ ciente	Alpha	238	201	Y	4.75087
+ aguardar	Alpha	268	198	Y	4.78056
+ reclamar	Alpha	217	183	Y	4.89422
+ cair	Alpha	201	179	Y	4.92611
+ constar	Alpha	205	175	Y	4.95871
+ verificar	Alpha	203	173	Y	4.97529
+ pagar	Alpha	198	168	Y	5.0176
+ entrar	Alpha	189	162	Y	5.07007
+ livrar	Alpha	192	160	Y	5.08799
ttt online	Unknown	176	151	Y	5.17152
+ ligar	Alpha	163	144	Y	5.24
+ pedir	Alpha	160	142	Y	5.26017
+ numerar	Alpha	171	138	Y	5.3014
+ enviar	Alpha	196	137	Y	5.31189
check list	Unknown	152	132	Y	5.36553
+ dever	Alpha	162	131	Y	5.3765
prox	Alpha	123	123	Y	5.46741
+ ponto ref prox	Unknown	114	114	Y	5.57703
+ back office disputar	Unknown	242	112	Y	5.60257
+ saldar	Alpha	131	112	Y	5.60257
+ ir	Alpha	123	111	Y	5.61551
+ saber	Alpha	121	107	Y	5.66845
prevencao consumo	Unknown	128	105	Y	5.69568

Figura 26: Atribuição de Pesos com o Método IDF

Na Figura 27 pode-se observar a disposição dos termos encontrados no dataset, em função de seu Peso e Frequência, após a atribuição de Pesos pelo Método IDF. Para efeito de exemplificação, foi identificado no gráfico, especificamente o termo "fraude", que obteve um peso de 0,39382 e apareceu nos diversos documentos analisados 187 vezes.

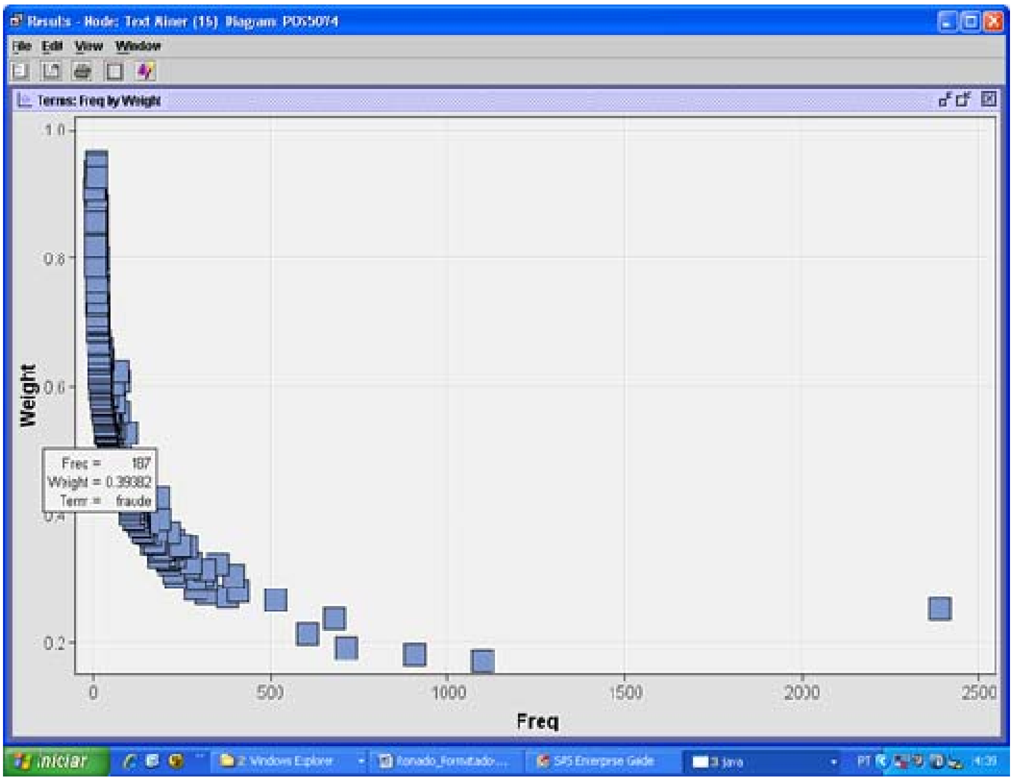
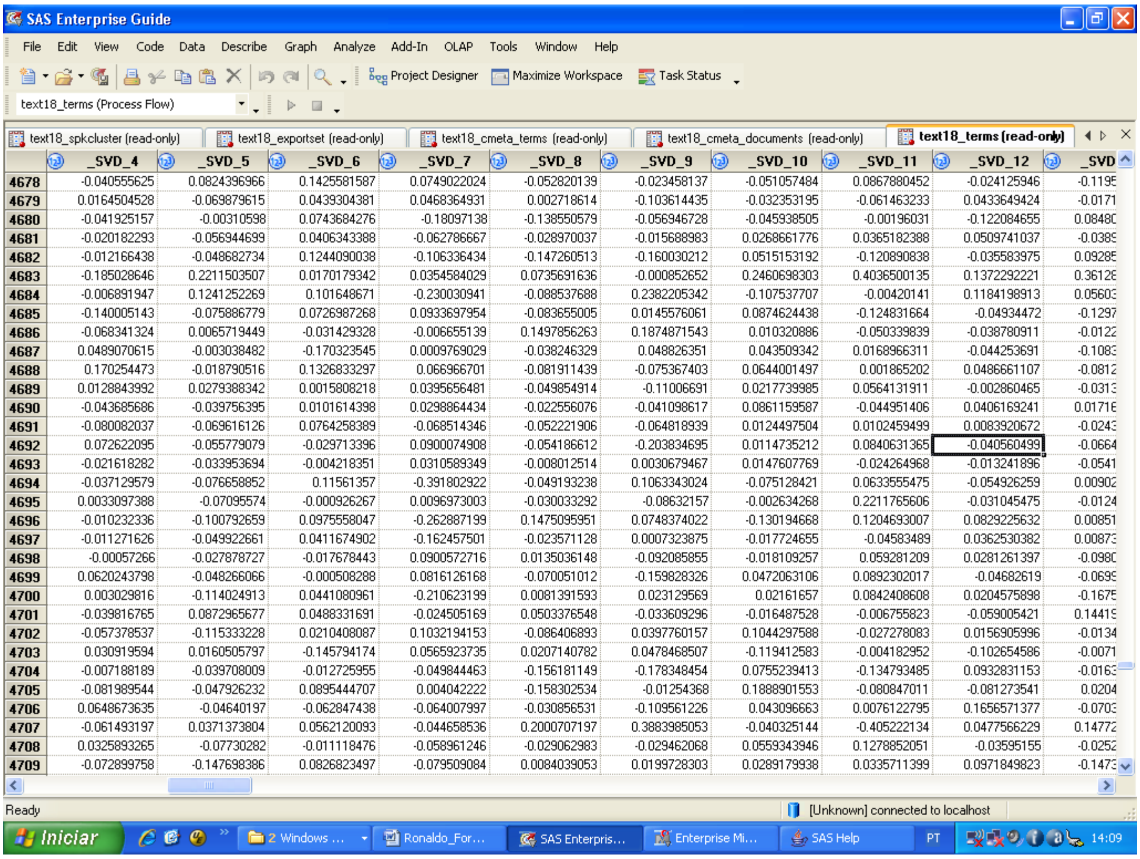


Figura 27: Pesos atribuídos a cada Termo

A Redução de Dimensionalidade

Para a redução de dimensionalidade foi utilizado o método de SVD (Singular Value Decomposition), mencionado na pesquisa bibliográfica (Capítulo 2). A partir da matriz original gerada para os 2517 documentos, foi gerada, com a utilização do SVD, uma matriz com um número bastante reduzido de termos, representando o dataset original.

A partir destes 2517 documentos, foram representados 5541 palavras. Por sua vez, os 5541 termos foram representados em uma matriz SVD de 100 termos, conforme Figura 28. Nesse Print Screen aparece apenas uma parte da matriz SVD gerada, dado que seria impossível visualizar toda a matriz em uma única página.



	_SVD_4	_SVD_5	_SVD_6	_SVD_7	_SVD_8	_SVD_9	_SVD_10	_SVD_11	_SVD_12	_SVD
4678	-0.040555625	0.0824396966	0.1425581587	0.0749022024	-0.052820139	-0.023458137	-0.051057484	0.0867880452	-0.024125946	-0.1195
4679	0.0164504528	-0.069879615	0.0439304381	0.0468364931	0.002718614	-0.103614435	-0.032353195	-0.061463233	0.0433649424	-0.0171
4680	-0.041925157	-0.00310598	0.0743684276	-0.18097138	-0.138550579	-0.056946728	-0.045938505	-0.00196031	-0.122084655	0.0848
4681	-0.020182293	-0.056944699	0.0406343388	-0.062786667	-0.028970037	-0.015688983	0.0268661776	0.0365182388	0.0509741037	-0.0389
4682	-0.012166438	-0.048682734	0.1244090038	-0.106336434	-0.147260513	-0.160030212	0.0515153192	-0.120890838	-0.035583975	0.09285
4683	-0.185028646	0.2211503507	0.0170179342	0.0354584029	0.0735691636	-0.000852652	0.2460698303	0.4036500135	0.1372292221	0.36128
4684	-0.006891947	0.1241252269	0.101648671	-0.230030941	-0.088537688	0.2382205342	-0.107537707	-0.00420141	0.1184198913	0.05603
4685	-0.140005143	-0.075886779	0.0726987268	0.0933697954	-0.083655005	0.0145576061	0.0874624438	-0.124831664	-0.04934472	-0.1297
4686	-0.068341324	0.0065719449	-0.031429328	-0.006655139	0.1497856263	0.1874871543	0.010320886	-0.050339839	-0.038780911	-0.0122
4687	0.0489070615	-0.003038482	-0.170323545	0.0009769029	-0.038246329	0.048826351	0.043509342	0.0168966311	-0.044253691	-0.1083
4688	0.170254473	-0.018790516	0.1326833297	0.066966701	-0.081911439	-0.075367403	0.0644001497	0.001865202	0.0486661107	-0.0812
4689	0.0128843992	0.0279388342	0.0015808218	0.0395656481	-0.049854914	-0.11006691	0.0217739985	0.0564131911	-0.002860465	-0.0313
4690	-0.043685686	-0.039756395	0.0101614398	0.0298864434	-0.022556076	-0.041098617	0.0861159587	-0.044951406	0.0406169241	0.01716
4691	-0.080082037	-0.069616126	0.0764258389	-0.068514346	-0.052221906	-0.064818939	0.0124497504	0.0102459499	0.0083920672	-0.0243
4692	0.072622095	-0.055779079	-0.029713396	0.0900074908	-0.054186612	-0.203834695	0.0114735212	0.0840631365	-0.040560499	-0.0664
4693	-0.021618282	-0.033953694	-0.004218351	0.0310589349	-0.008012514	0.0030679467	0.0147607769	-0.024264968	-0.013241896	-0.0541
4694	-0.037129579	-0.076658852	0.11561357	-0.391802922	-0.049193238	0.1063343024	-0.075128421	0.0633555475	-0.054926259	0.00902
4695	0.0033097388	-0.07095574	-0.000926267	0.0096973003	-0.030033292	-0.08632157	-0.002634268	0.2211765606	-0.031045475	-0.0124
4696	-0.010232336	-0.100792659	0.0975558047	-0.262887199	0.1475095951	0.0748374022	-0.130194668	0.1204693007	0.0829225632	0.00851
4697	-0.011271626	-0.049922661	0.0411674902	-0.162457501	-0.023571128	0.0007323875	-0.017724655	-0.04583489	0.0362530382	0.00873
4698	-0.00057266	-0.027878727	-0.017678443	0.0900572716	0.0135036148	-0.092085855	-0.018109257	0.059281209	0.0281261397	-0.0980
4699	0.0620243798	-0.048266066	-0.000508288	0.0816126168	-0.070051012	-0.159828326	0.0472063106	0.0892302017	-0.04682619	-0.0699
4700	0.003029816	-0.114024913	0.0441080961	-0.210623199	0.0081391593	0.023129569	0.02161657	0.0842408608	0.0204575898	-0.1675
4701	-0.039816765	0.0872965677	0.0488331691	-0.024505169	0.0503376548	-0.033609296	-0.016487528	-0.006755823	-0.059005421	0.14419
4702	-0.057378537	-0.115333228	0.0210408087	0.1032194153	-0.086406893	0.0397760157	0.1044297588	-0.027278083	0.0156905996	-0.0134
4703	0.030919594	0.0160505797	-0.145794174	0.0565923735	0.0207140782	0.0478468507	-0.119412583	-0.004182952	-0.102654586	-0.0071
4704	-0.007188189	-0.039708009	-0.012725955	-0.049844463	-0.156181149	-0.178348454	0.0755239413	-0.134793485	0.0932831153	-0.0163
4705	-0.081989544	-0.047926232	0.0895444707	0.004042222	-0.158302534	-0.01254368	0.1888901553	-0.080847011	-0.081273541	0.0204
4706	0.0648673635	-0.04640197	-0.062847438	-0.064007997	-0.030856531	-0.109561226	0.043096663	0.0076122795	0.1656571377	-0.0703
4707	-0.061493197	0.0371373804	0.0562120093	-0.044658536	0.2000707197	0.3883985053	-0.040325144	-0.405222134	0.0477566229	0.14772
4708	0.0325893265	-0.07730282	-0.011118476	-0.058961246	-0.029062983	-0.029462068	0.0559343946	0.1278852051	-0.03595155	-0.0252
4709	-0.072899758	-0.147698386	0.0826823497	-0.079509084	0.0084039053	0.0199728303	0.0289179938	0.0335711399	0.0971849823	-0.1473

Figura 28: Amostra da Matriz SVD gerada

A quantidade ideal, levando-se em consideração a *acurácia* da representatividade da matriz original e o tempo de processamento, foi obtida de maneira interativa, ou seja, foram feitos testes com matrizes menores e maiores. A melhor representação foi obtida com 100 termos (k=100)..

5.3 O PROCESSAMENTO DOS TEXTOS

Foram feitas várias tentativas de Clusterização. A melhor representação obtida foi com um total de 5 Clusters, que serão detalhados na sequência. Estes clusters foram obtidos através do Método de Expectation Maximizaton.

Apenas para evidenciação do processo interativo de obtenção dos Clusters Finais, observa-se nas Figuras 29 e 30 a distribuição de quantidade de documentos e localização dos Centróides para as primeiras tentativas de Clusterização. As primeiras tentativas de agrupamento foram realizadas com 10 agrupamentos.

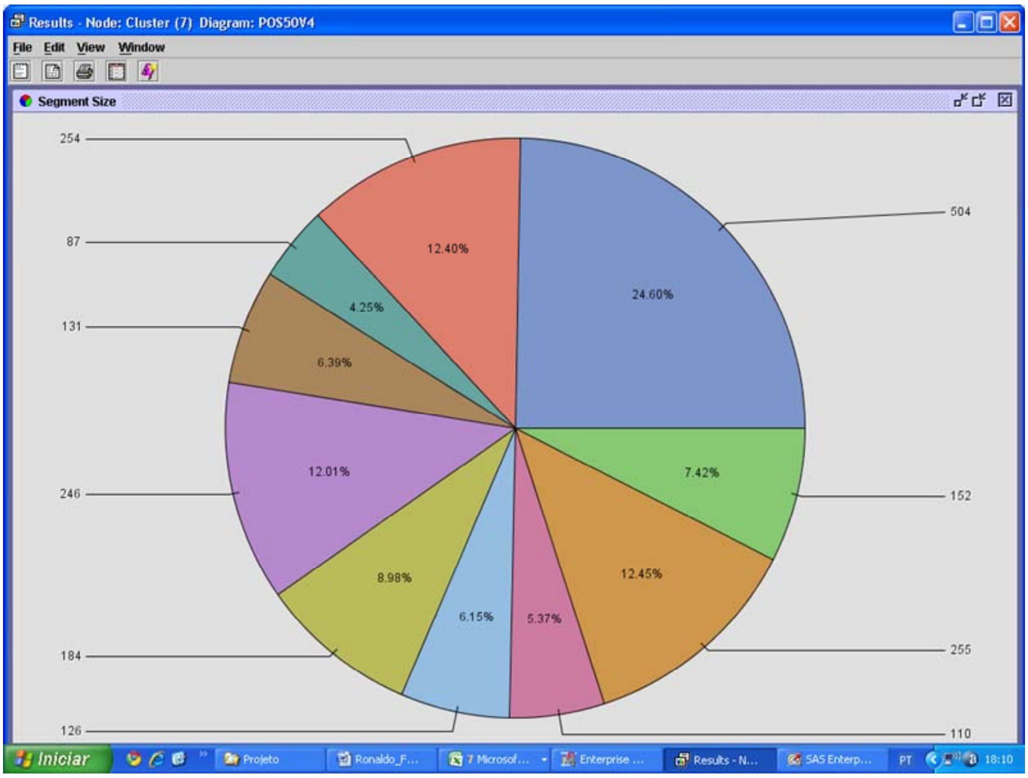


Figura 29: Distribuição da qtd. documentos por Cluster (na Clusterização Inicial)

Na Figura 30, e também na Tabela 6, evidencia-se a necessidade de um maior agrupamento dos clusters iniciais, dado que é possível observar que os Centróides estão muito próximos uns dos outros.

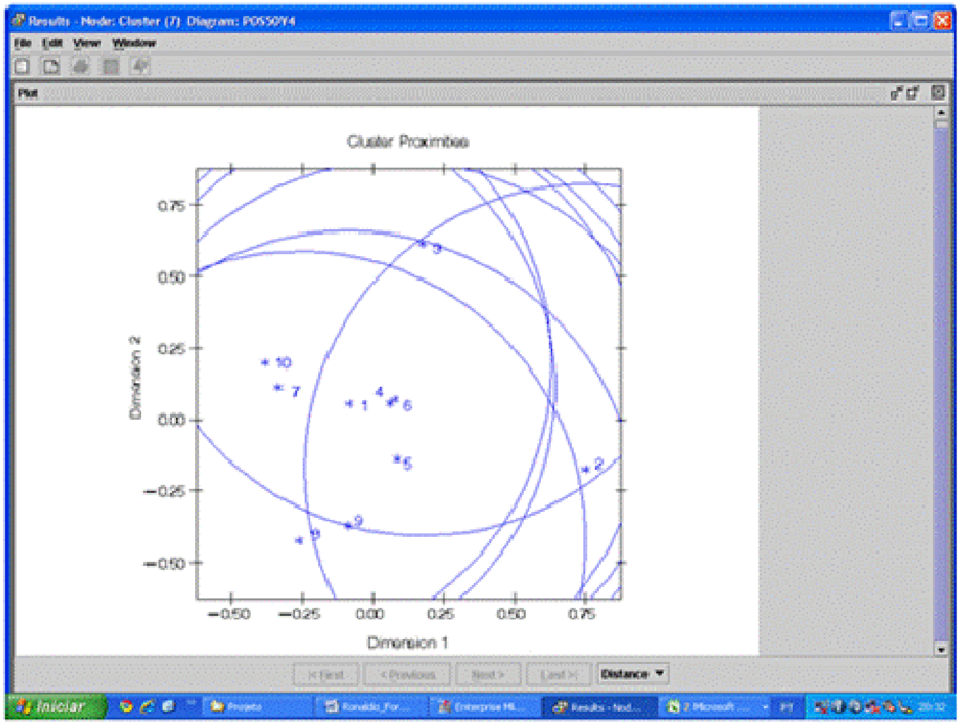


Figura 30: Localização Relativa dos Centróides (processo inicial de Clusterização)

Na Tabela 6 os valores correspondentes aos Centróides evidenciados na Figura 35:

Segment	Frequency of Cluster	Root Mean Square Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
1	504	0,09253762	1,04447746	5	0,3677086
2	254	0,05245734	0,99422602	1	0,8454878
3	87	0,07197664	1,02019345	1	0,6982485
4	131	0,07816462	0,98421344	1	0,589484
5	246	0,09060257	1,04024327	1	0,3677086
6	184	0,08875519	1,03120579	1	0,3844477
7	126	0,08753921	0,9784011	1	0,4562972
8	110	0,07689351	1,01194441	1	0,5949739
9	255	0,07837734	1,02842253	1	0,5773063
10	152	0,0840084	1,01169594	1	0,4780614

Tabela 6: Estatísticas dos Primeiros Clusters encontrados

A seguir a descrição das características dos cinco clusters finais obtidos após o refinamento do processo interativo utilizando o algoritmo de *Expectation Maximization*:

No Anexo Figuras a seguir as características estatísticas dos cinco clusters (clusters finais obtidos):

Na separação considerada como a melhor possível, observa-se uma boa separação entre o Centróides, conforme Figura 31.

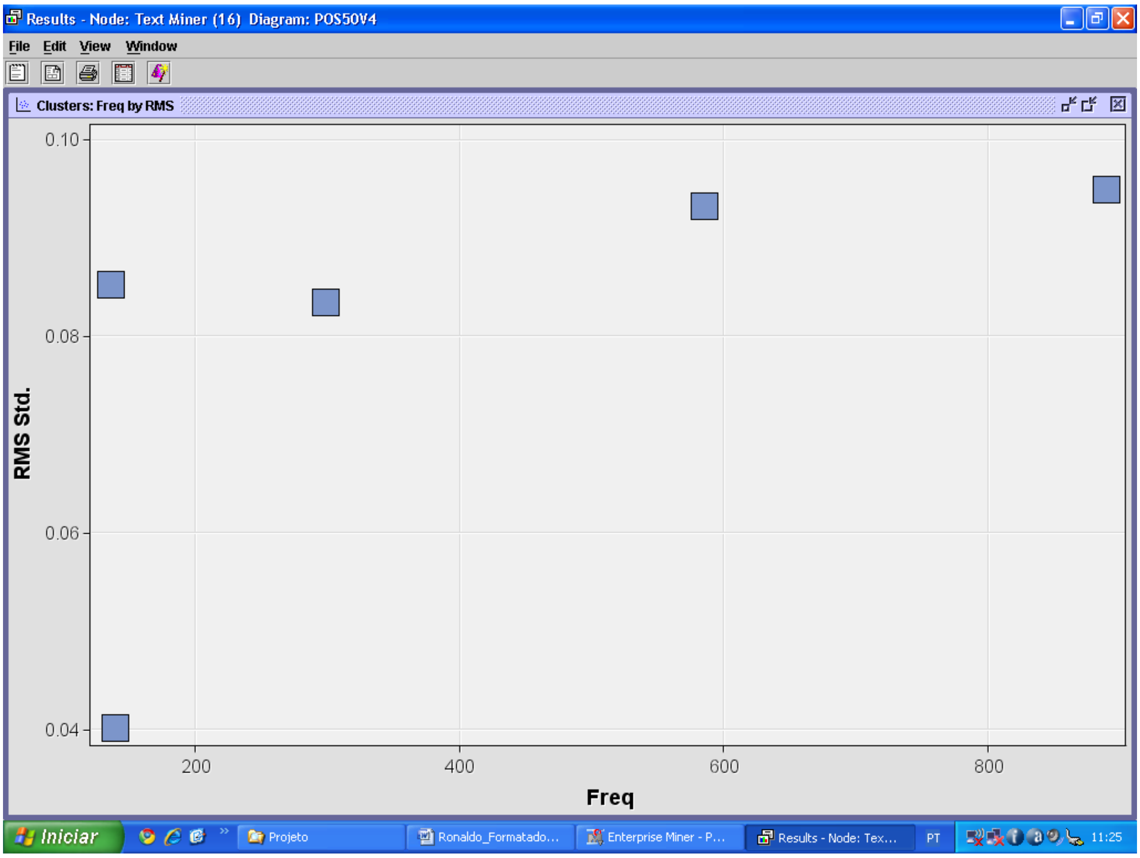


Figura 31: Localização dos Centróides dos Clusters Finais

Conceitos Extraídos dos Clusters Obtidos

Como foi dito, após diversas interações e com a exclusão de muitos termos que ocorriam em vários clusters, esta classificação com cinco agrupamentos foi a que fez mais sentido, através da observação do conjunto de textos e pela análise de correlação (de termos) evidenciada a seguir.

Na Tabela 7, observa-se a relação de Palavras-Chave que melhor caracterizam os cinco clusters encontrados:

Cluster	Palavras-Chave
1	efetuar, confirmar, antigo, novo, contato, sucesso, atendido, mover, email
2	ponto, final, entrega, depositar, ponto referência próximo, identificar, reparo aberto, visar, ponto referência próximo
3	solicitar, email, vencimento, enviado, fatura, codigo, envio, pedido
4	mover, saldo, feito susp, lig caiu, atendimento cs livre, informar
5	plano, informar, aparelho, fazer, atendimento, aberto, fazer, desejar, orientar, linha

Tabela 7: Palavras-Chave dos Clusters Finais

5.3.1 Descrição / Resumo dos Clusters Obtidos

Cluster 1: Problemas relacionados a Reparos

Cluster 2: Problema de Logística de Entrega (Aparelho e Fatura de Clientes)

Cluster 3: Solicitação de Segunda Via / Código de Barras

Cluster 4: Solicitações de Migração de Saldo de um Ciclo (de Pagamento para Outro)

Cluster 5: Questionamentos sobre o Valor da Conta

A seguir um maior detalhamento de cada um dos Clusters mencionados, com os respectivos mapas de inter-relações entre os principais termos mais representativos ((indicados na Tabela 7).

<u>Cluster 1: Reparos</u>

Neste agrupamento ao verificar-se a coleção de textos em conjunto e ao verificar-se a relação entre as palavras de maior importância, observa-se-se claramente como idéia preponderante as solicitações de reparo solicitadas por clientes (em suas respectivas linhas telefônicas ou em seus aparelhos). Nas tela de análise de correlação entre termos deste *cluster* (Figura 32), observa-se que as relações mais fortes aparecem com linhas mais grossas e as menos intensas com linhas mais finas.

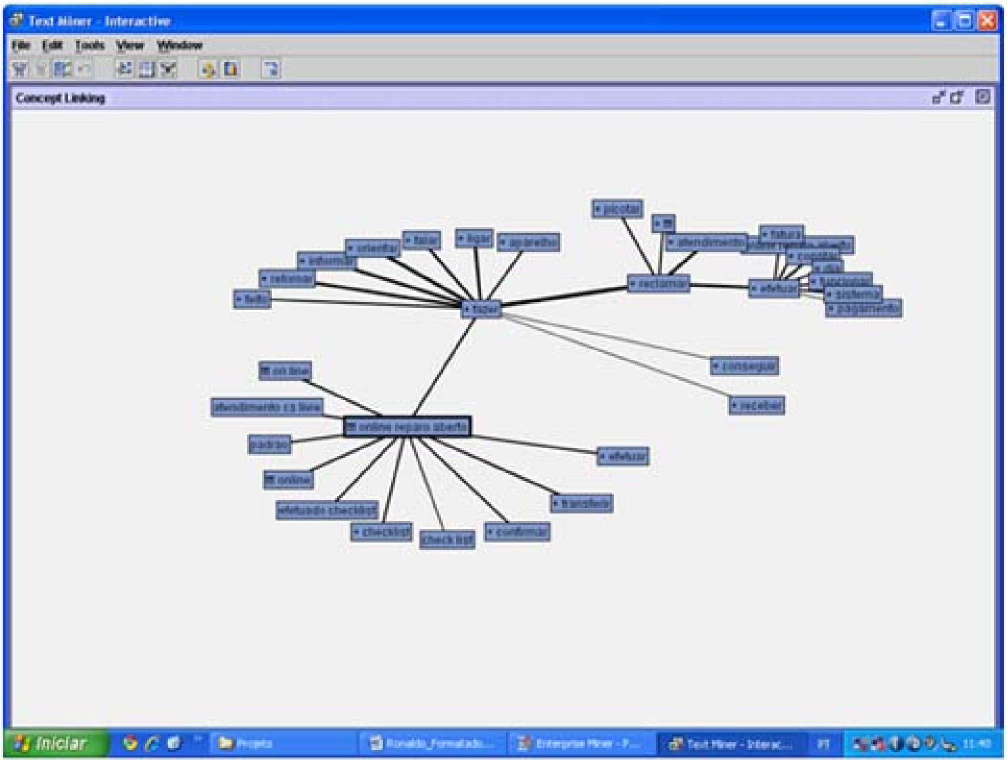


Figura 32: Correlação entre os Termos Mais Explicativos do Cluster 1

Na Transcrição 1 (Apêndice E), observa-se um texto já classificado no Cluster 1, para exemplificação. Os dados pessoais do Cliente foram omitidos propositalmente (marcados com “xxx”):

Vale ressaltar, a importância da etapa de Pré-Processamento, neste Estudo de Caso, dada a grande quantidade de abreviações utilizada pelos Operadores que realizaram o Atendimento e por sua vez, escreveram cada um dos textos analisados. Por esse motivo foi necessária uma atenção especial à etapa de Pré-Processamento, com muitas interações e aperfeiçoamento contínuo do *parsing* - de forma a não inviabilizar a etapa de extração de conhecimento e conclusões. Desta forma foi necessário analisar e adaptar os algoritmos a uma quase *Pseudo-Linguagem,* adaptada aos jargões e terminologias técnicas dos Operadores de Call Center.

Cluster 2: Problema de Logística de Entrega (Aparelho e Fatura de Clientes)

Na Figura 33 observa-se análise de correlação entre termos representativos do Cluster 2.

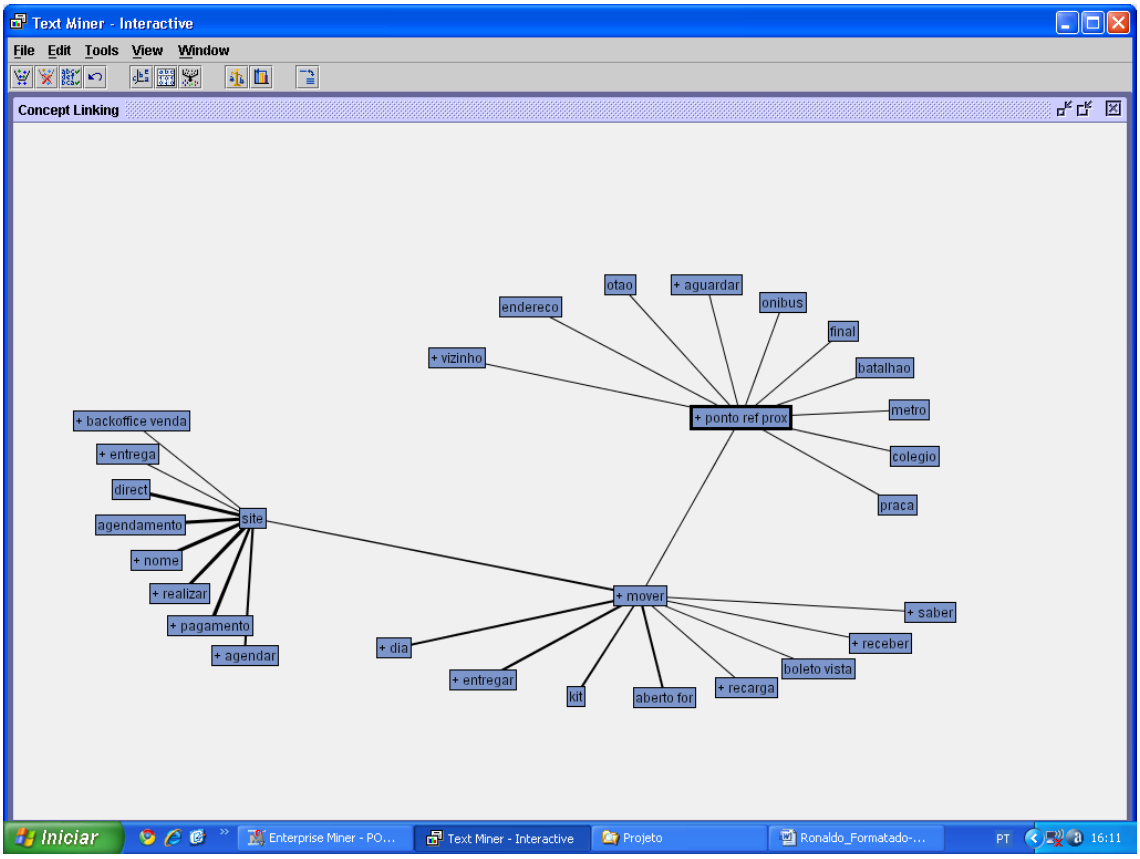


Figura 33: Correlação entre os Termos Mais Explicativos do Cluster 2

No Apêndice E, observa-se na Transcrição 2, um texto classificado do Cluster 2 . A partir da análise das correlações entre os termos e a conseqüente validação (destas correlações) com a transcrição dos textos já classificados, realmente pode-se afirmar tratar-se de um conjunto (de textos) relacionados a problemas de Logística de Entrega

Cluster 3: Solicitação de Segunda Via / Boleto com Código de Barras para Pagamento

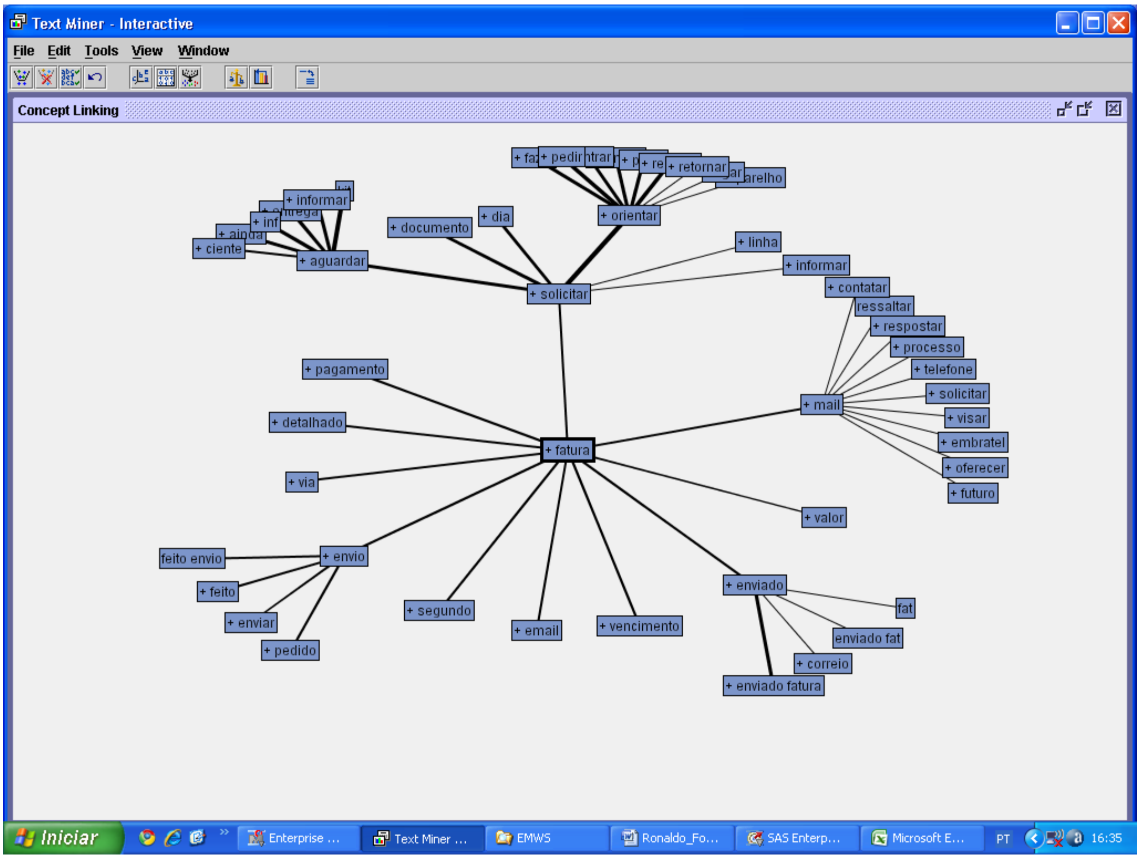


Figura 34: Correlação entre os Termos Mais Explicativos do Cluster 3

A partir da avaliação das correlações entre os termos observados na Figura 34 e das Transcrições (exemplificadas nas Transcrições 3 e 4 do Apêndice E), fica nítida que o Cluster 3 está diretamente relacionado a questões de Solicitações de Segunda Via de Contas e / ou Boletos com Códigos de Barras para pagamentos.

Cluster 4: Solicitações de Migração de Saldo de um Ciclo (de Pagamento para Outro)

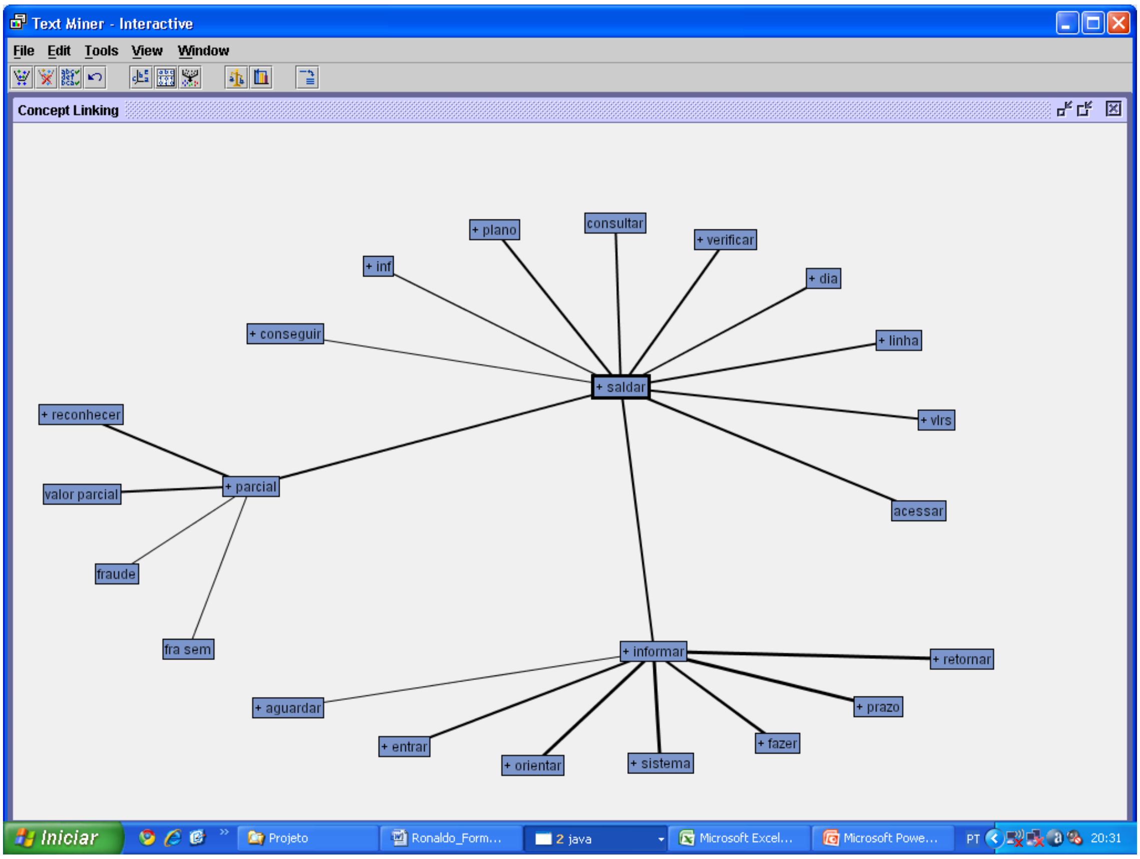


Figura 35: Correlação entre os Termos Mais Explicativos do Cluster 4

A partir da avaliação das correlações entre os termos observados na Figura 35 e Transcrições correspondentes (Apêndice E), verifica-se que "conceito" de Solicitações de Migração de Saldo de Contas, como predominante.

Além disso, é interessante observar que, na Figura 35, o termo "saldar" aparece com muitas correlações relevantes. Contudo ao verificar na lista de termos, que inclui os radicais e suas respectivas variações, ou seja, no agrupamento realizado pelo processo de stemming, verificamos que na verdade o termo que aparece com maior freqüência é o termo "saldo" (conforme Figura 36). O termo "saldo" refere-se a um "saldo de conta" e desta forma vai de encontro com a idéia central do conjunto de Textos agrupados no Cluster 4, ou seja, a migração de Saldos de Conta.

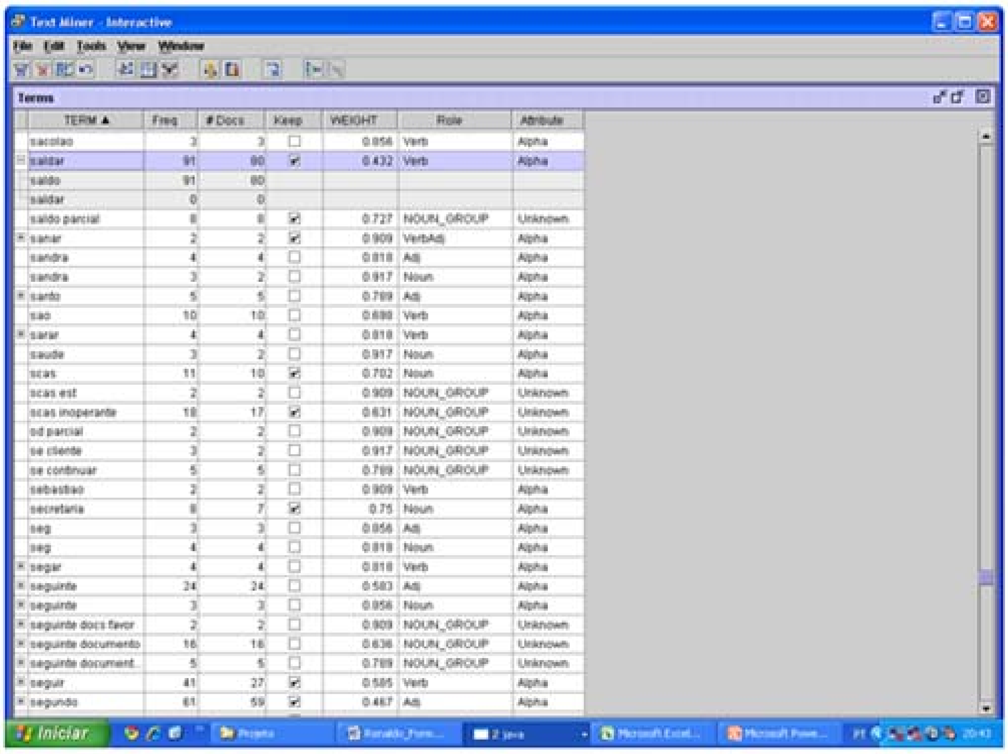


Figura 36: Verificação do Termo e seu respectivo Radical

Cluster 5: Questionamentos sobre o Valor da Conta

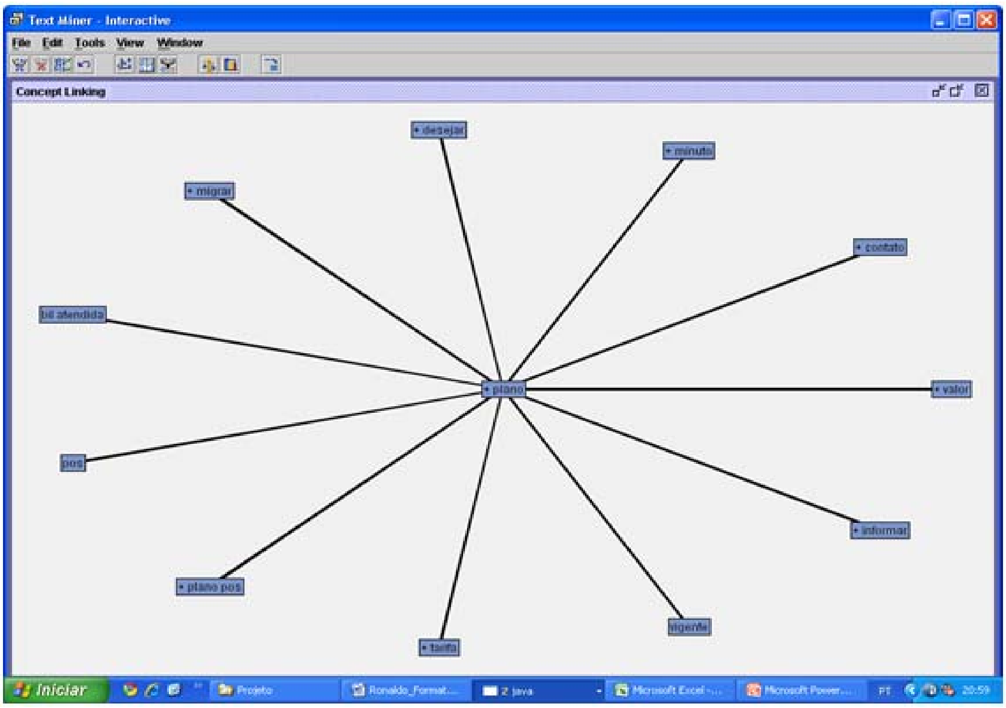


Figura 37: Correlação entre os Termos Mais Explicativos do Cluster 5

A análise das correlações entre os termos mais representativos dos Cluster 5 (Figura 37) deixa algumas margens de dúvidas sobre suas características principais. Estas dúvidas são dissipadas quando as Transcrições dos Textos classificados são analisadas – deixando claro tratarem-se de textos relacionados a questionamentos de valor de conta. No Apêndice E, observa-se a Transcrição 6 exemplificando isto.

5.4 O PROCESSO DE GEOREFERENCIAMENTO

Estando cada um dos textos já atrelado a um único Cluster (etapa de Mineração de Textos já realizada) e a um respectivo número de telefone – característica da base de dados obtida, pode-se agora executar a complementação da análise através da visão do Georeferenciamento, possibilitando desta forma "enxergar" as solicitações dos Clientes sob um novo prisma, ou seja, observando como é a distribuição geográfica destes Clusters no Espaco Geográfico e seus respectivos desdobramentos sócio-demográficos.

Neste trabalho, o SIG utilizado foi o MapInfo e forma de representação foi "Vetorial", conforme item 4.2.1 (da Metodologia). Este SIG organiza seus dados em forma de tabelas. Cada tabela é um grupo de arquivos que constitui um mapa ou de banco de dados. Estes arquivos estão associados uns aos outros de forma que, para a leitura e manipulação deste banco, é necessário que todos os arquivos Mapinfo estejam disponíveis em um mesmo local físico (mesmo diretório)

Os arquivos Mapinfo e suas respectivas funcionalidades podem ser descritos como:

***.tab**: descreve a estrutura da tabela, a organização e formato dos dados tabulares

***.dat**: contém os dados tabulares, o conteúdo de cada tabela. Caso vocês esteja trabalhando com arquivos provindos de outros softwares, sua tabela será composta pela junção do arquivo *.dat com o arquivo de origem, que pode ser *.xls, *.dbf, *tif, *.jpg, entre outros.

***.map**: descreve objetos gráficos

***.id**: arquivo que vincula os objetos gráficos (*.map) aos dados tabulares (*.dat).

5.4.1 A Representação da Área de Cobertura

As bases de dados para Georeferenciamento, foram obtidas da Operadora de Telecomunicações e de fonte pública (IBGE), especificamente para o caso de informações referentes a Classes Sociais. Por questões de Segurança de Informação e de confidencialidade da Área de Cobertura da Operadora, as novas ERBs da área de cobertura (situação atual) não foram *plotadas*. A Área de Cobertura descrita, representa a situação no período de janeiro a setembro de 2008, na Região Metropolitana do Rio de Janeiro, ou seja, no período em que os registros testuais foram coletados.

Para a representação da Área de Cobertura da Operadora e seus respectivos entornos geográficos, foram utilizadas várias camadas (“layers”) superpostas, descritas a seguir:

- layer de bairros

- layer de ruas

- layer com as delimitações das Estações Radio-Base

A Representação das Camadas

Para inserir as camadas descritas no item anterior, foi utilizado o acoplamento de tabelas, conforme metodologia (item 4.2.1 / sub-item 21.1). Desta forma, conforme exemplificação nas Figuras 38 e 39, no Menu Tabela / Atualizar Colunas, seleciona-se a camada de referência e a respectiva coluna desta que será atualizada (neste caso foi propositalmente deixada uma coluna apenas com seu respectivo *label*) com os novos valores, referentes às coordenadas x,y dos documentos. Importante que nesta forma de “casamento” de bases exista uma chave comum.

Figura 38: Passo 1de 2 para atualização de Tabela

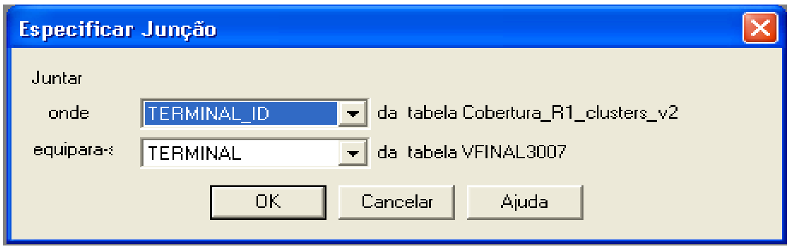


Figura 39: Passo 2 de 2 para atualização de tabela

Para a execução dos mapas temáticos dos clusters no SIG foram percorridos os seguintes passos:

- Escolha entre os seguintes tipos de mapas: manchas, gráfico de barras, sectogramas, tamanho, textura, forma e grelha. Deve-se escolher o tipo que considerado com a melhor representação da variável temática

- Escolha da tabela e respectiva coluna (desta) a ser representada tematicamente. Isto pode ser feito para uma coluna já existente ou também a partir de uma função matemática que utilize os valores de quaisquer colunas da tabela de dados de uma determinada camada (*layer*)

- Divisão de faixas conforme valores pré-definidos ou (como neste trabalho) utilizar separação customizada de classes. Neste caso cada classe refere-se a valores relacionados a maior Lou menor incidência do cluster (de documentos) que estiver sendo representado. Nesta etapa também são definidos cores e tramas do mapa, assim como as legendas relacionadas.

A seguir a visão de algumas destas camada vistas separadamente, após o “casamento” das tabelas, facilitando o entendimento para as explanações subseqüentes (Figuras, 40, 41,42, 43 e 44):

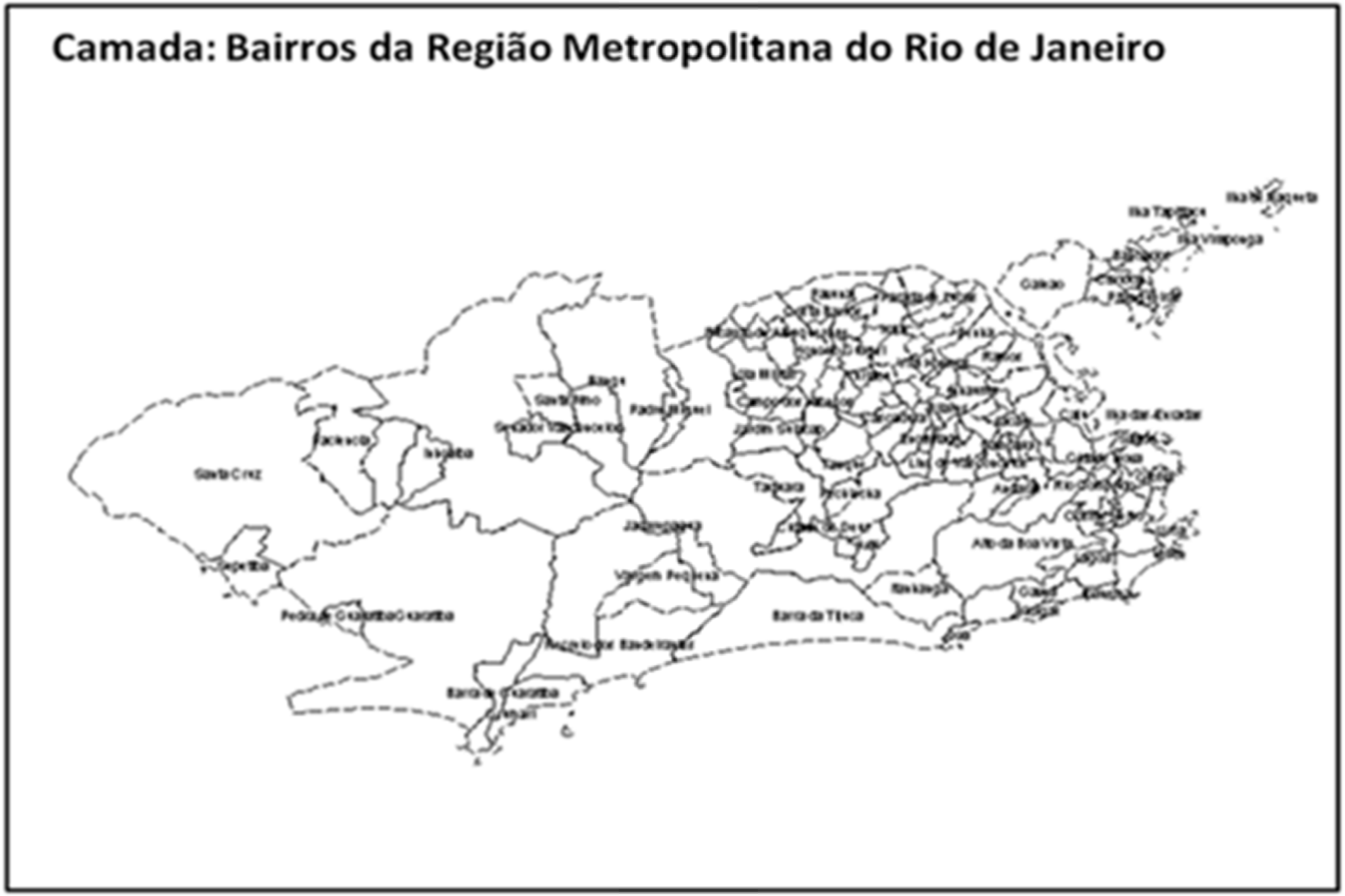


Figura 40

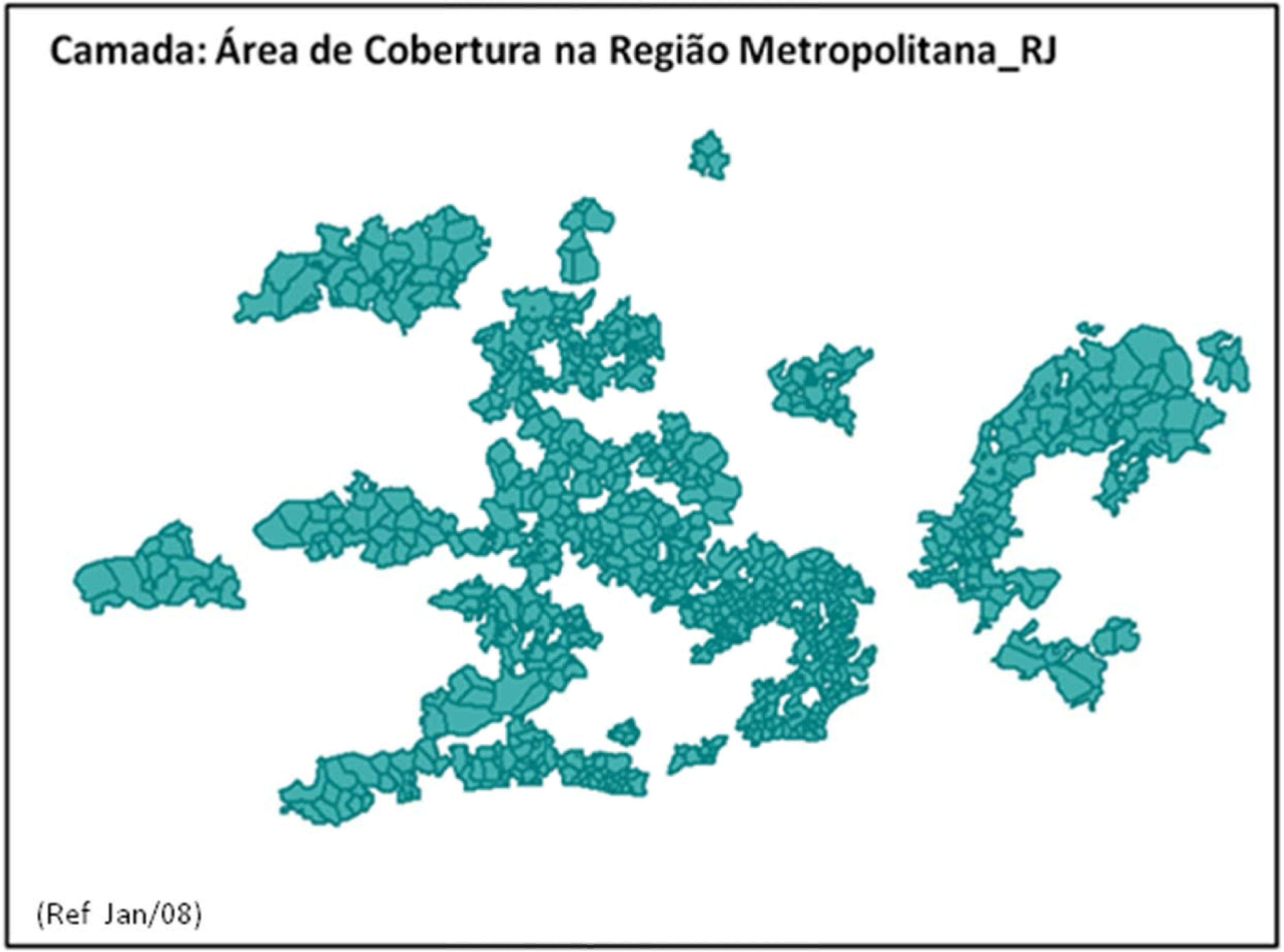


Figura 41: Área de Cobertura

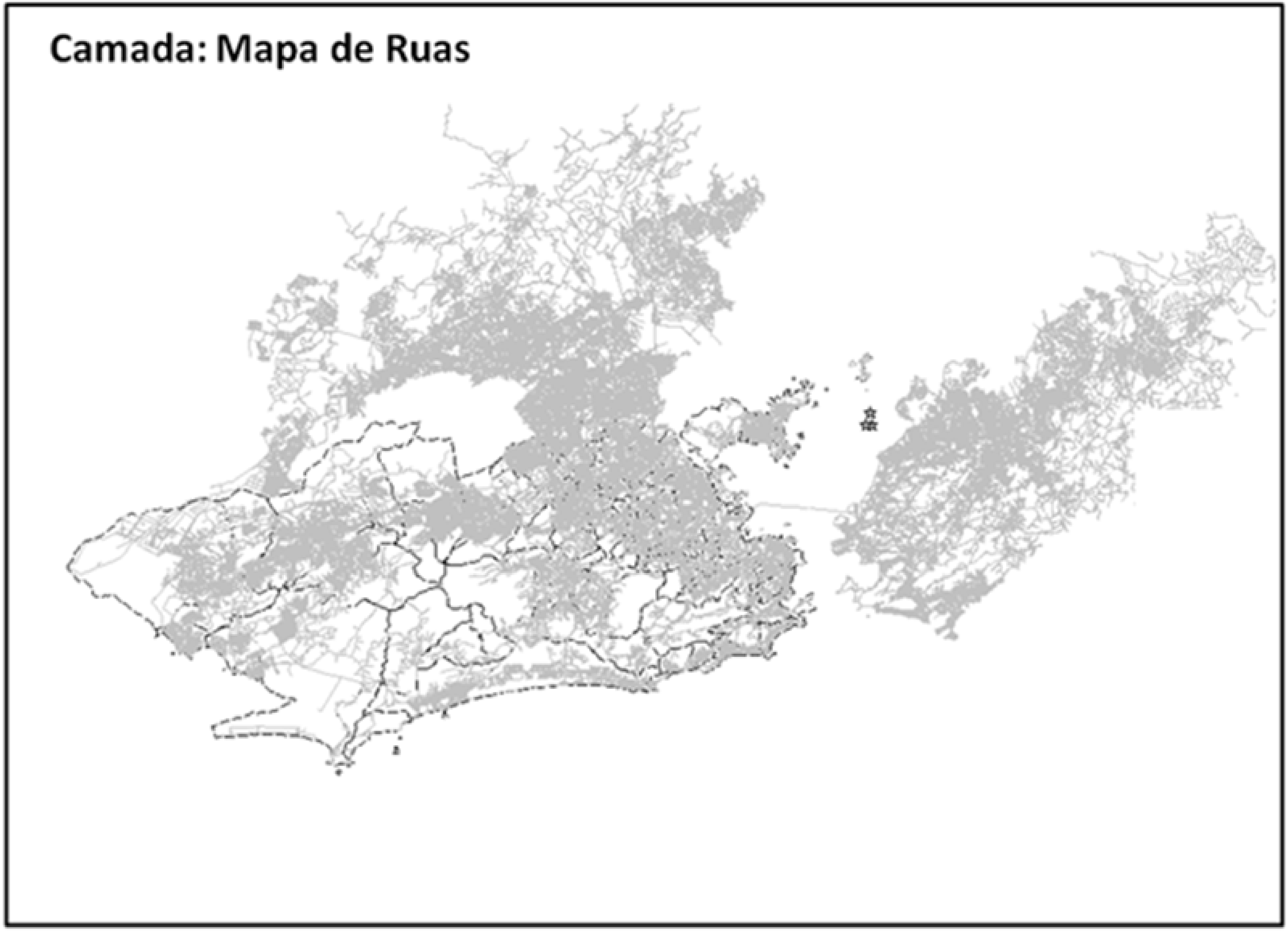


Figura 42: Mapa de Ruas

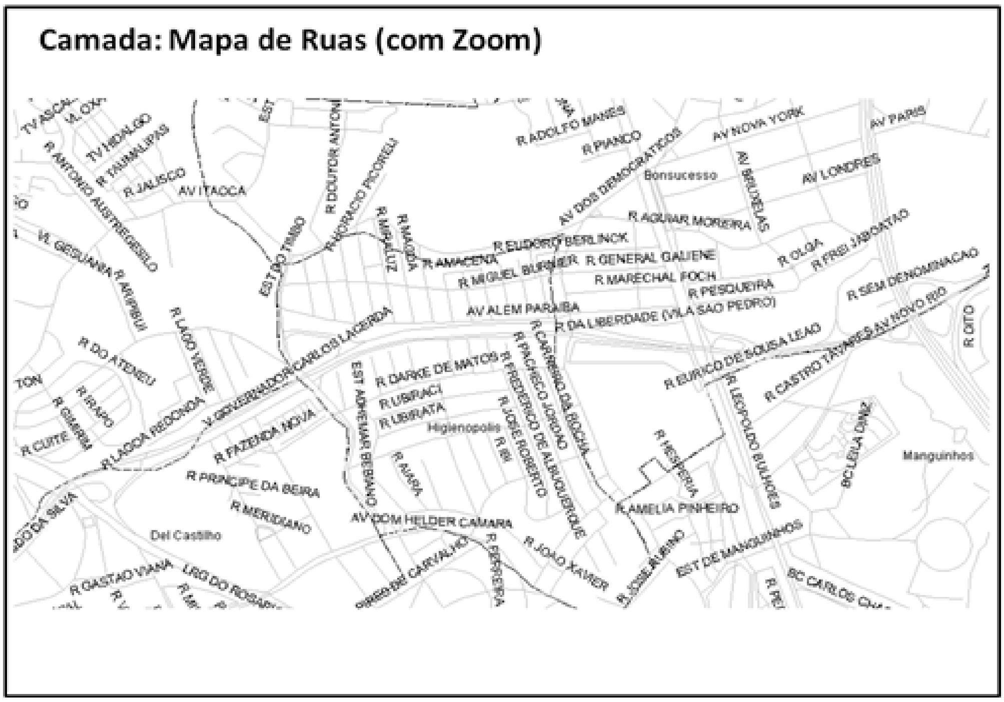


Figura. 43: Representação da Camada de Ruas com *"zoom"*

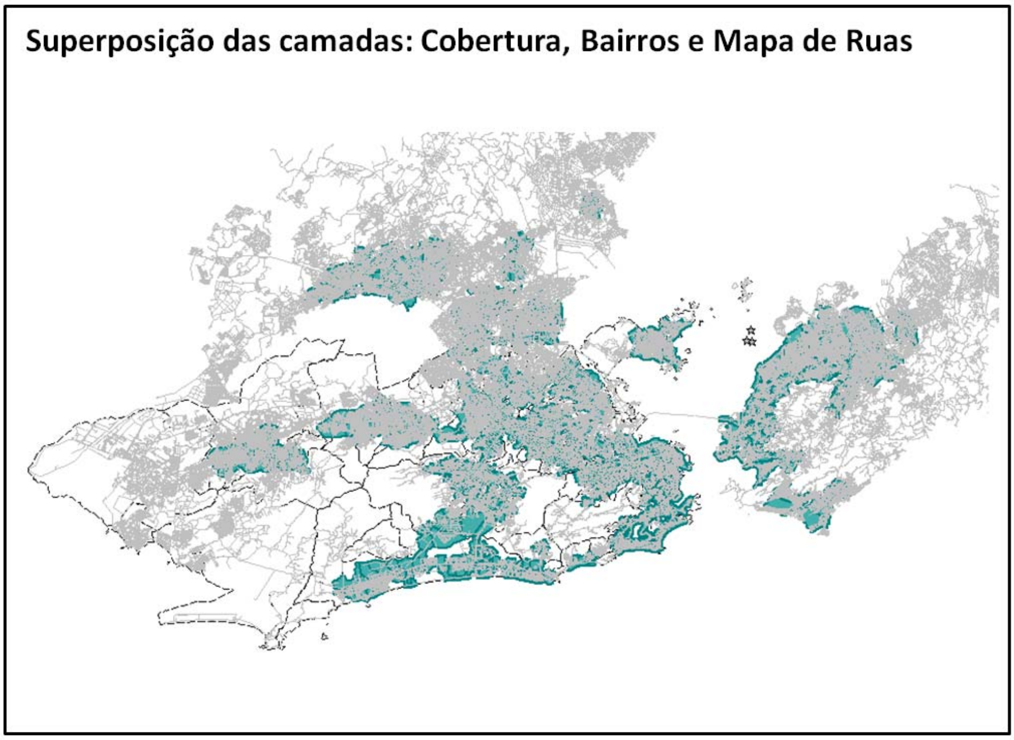


Figura 44: Representação da Superposição de Camadas

A Camada “Classe Social”

Uma importante etapa, que é útil na análise da representação dos Clusters é o Georeferenciamento das Classes Sociais na área coberta pelo Estudo.

Para a representação das Classes Sociais, utilizaram-se dados internos da Operadora que gentilmente cedeu os dados para o Estudo de Caso. Estas classificações foram baseadas por sua vez em informações do IBGE. Ou seja, para cada Estação Radio Base (ERB), foi possível totalizar a quantidade respectiva de Clientes das Classes A, B, C, D e E. No mapa da Figura 45, o critério foi totalizar a quantidade de terminais (por ERB) das Classes A e B e dividi-las pela quantidade total de terminais (de todas as Classes), ou seja, de forma “relativizada”. Desta forma, é possível obter um mapa temático onde percebe-se a maior ou menor incidência de consumidores com alta renda.

Observa-se na Figura 45 incidência maior de concentração de alta renda em regiões da Zona Sul e particularmente nas ERBs que cobrem a orla da Barra da Tijuca. Como ressalva observa-se que estes dados (de Classe Social) foram baseados em estudo realizado há cerca de dois anos atrás, em função da publicação da última pesquisa do gênero do IBGE. Além disso, em média, cerca de 25% dos terminais não estão classificados, ou seja, o mapa foi baseado na classificação dos outros 75% de terminais que possuíam classificação. Ainda assim, tal representação, em conjunto com as demais que análises ajuda o entendimento do comportamento dos consumidores avaliados no Estudo de Caso.

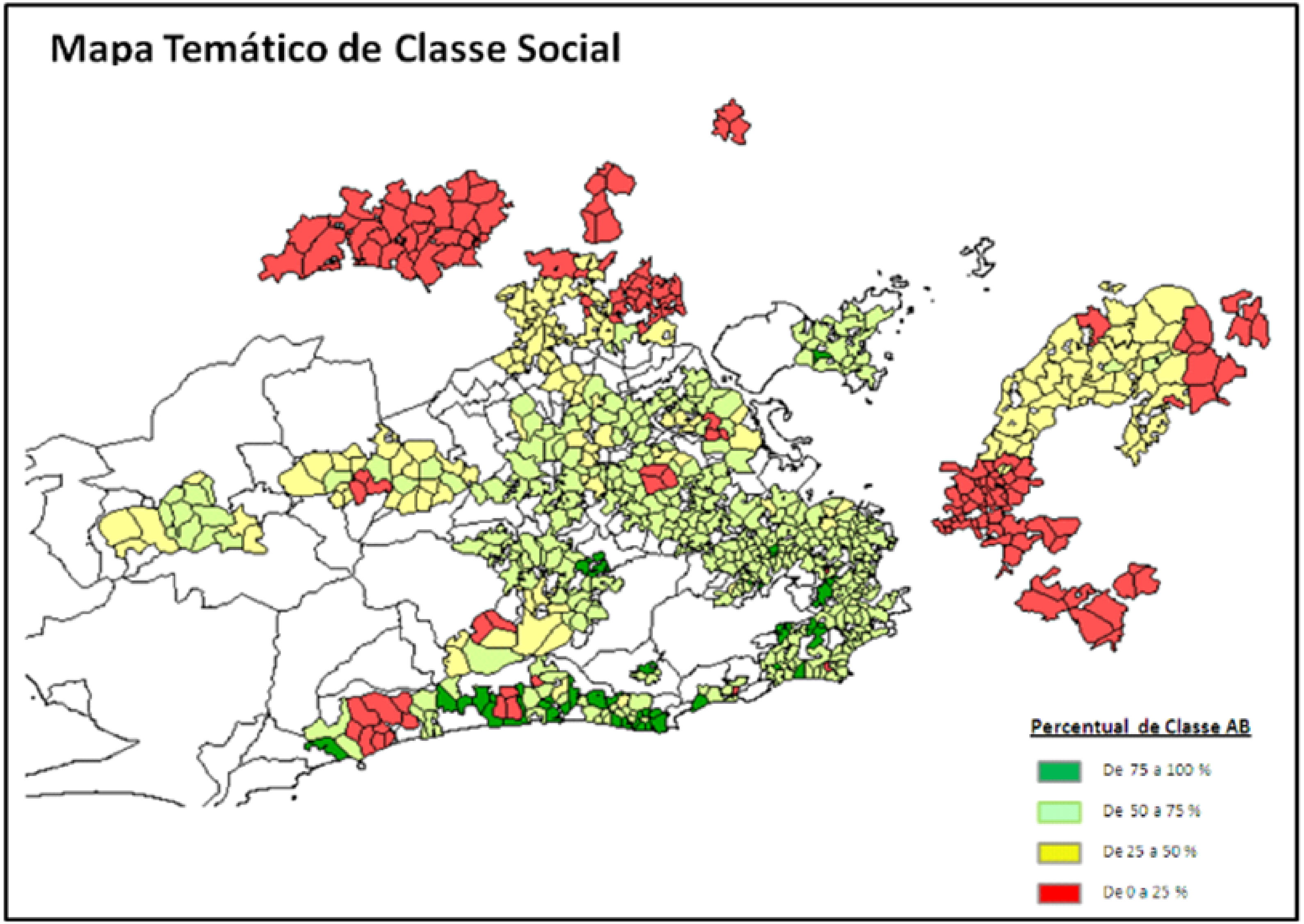


Figura 45: Representação de Classes Sociais

Apenas para corroborar a validade dos dados de identificação de ERBs por Classe Social, observa-se na Figura 46, um "*Zoom*" da região de Ipanema, onde sabe-se existir grande concentração de Classes AB, conforme evidenciado no mapa de Classe Social (Figura 46). Entretanto, observa-se também uma área muito próxima, cuja concentração de Classe AB é diametralmente oposta, que é a Comunidade do Morro do Cantagalo, evidenciada com a cor vermelha no mapa a seguir.

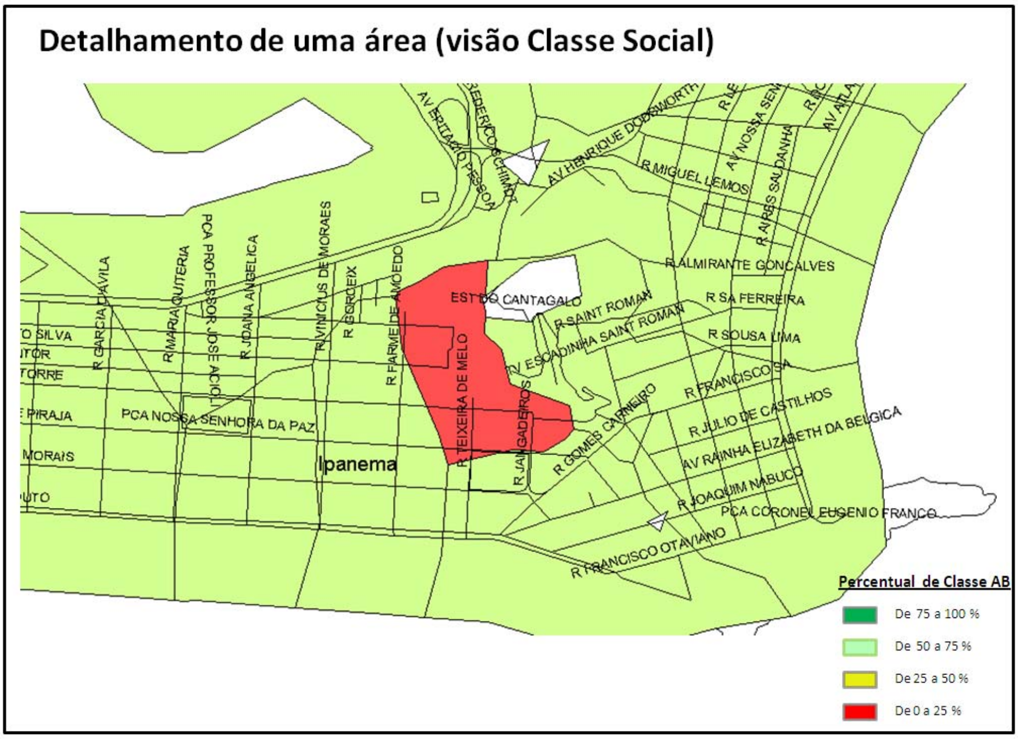


Figura 46: Detalhamento ("Zoom") de uma área da Zona Sul do Rio de Janeiro

O Georeferenciamento dos Clusters

A partir do Georeferenciamento de cada um dos Terminais de Clientes, é possível identificar exatamente a qual ERB estes pertencem. Tendo esta informação é possível criar um Mapa Temático no GIS, identificando através de cores específicas as concentrações de Clientes de cada Cluster em cada uma das ERBs. O problema que decorreu desta sistemática foi o fato de existir a possibilidade de em um local de grande incidência de solicitações para o Call Center estar em uma área em que a densidade (quantidade de clientes por área geográfica) de Clientes ser grande. A solução encontrada para isto foi relativizar as informações temáticas do mapa, ou seja, as identificações passaram a ser quantidade de ligações (ou reclamações) divididas pela quantidade de clientes, este número variando agora entre 0 e 1, pois não houve qualquer ERB onde tenha ocorrido um número de solicitações / reclamações maior que a quantidade de Assinantes. Nas figuras a seguir as representações dos Clusters com a utilização desta metodologia (Figuras 47, 48, 49, 50 e 51):

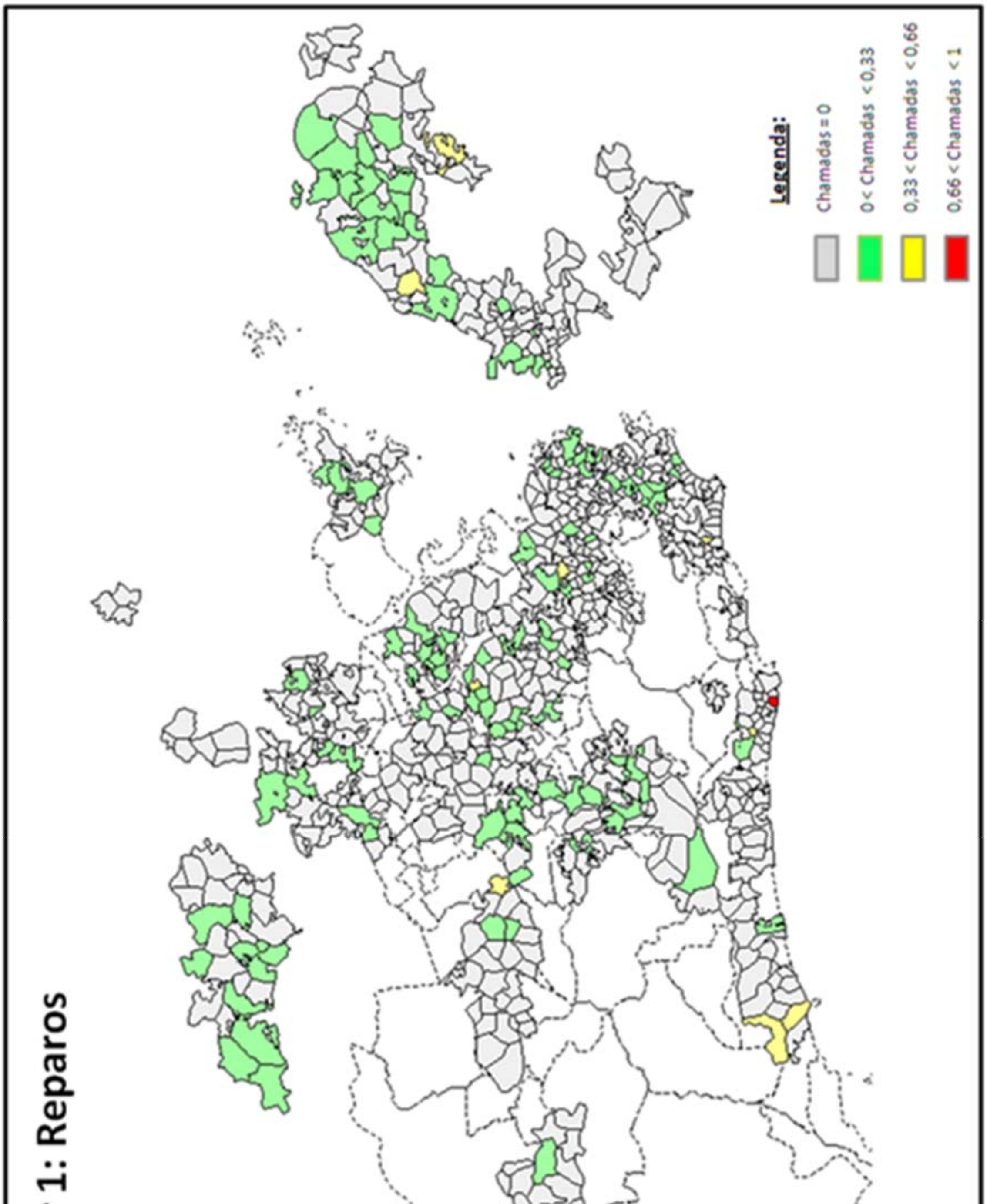

Cluster 1: Reparos
Legenda:
Chamadas = 0
0 < Chamadas < 0,33
0,33 < Chamadas < 0,66
0,66 < Chamadas < 1

Figura 47

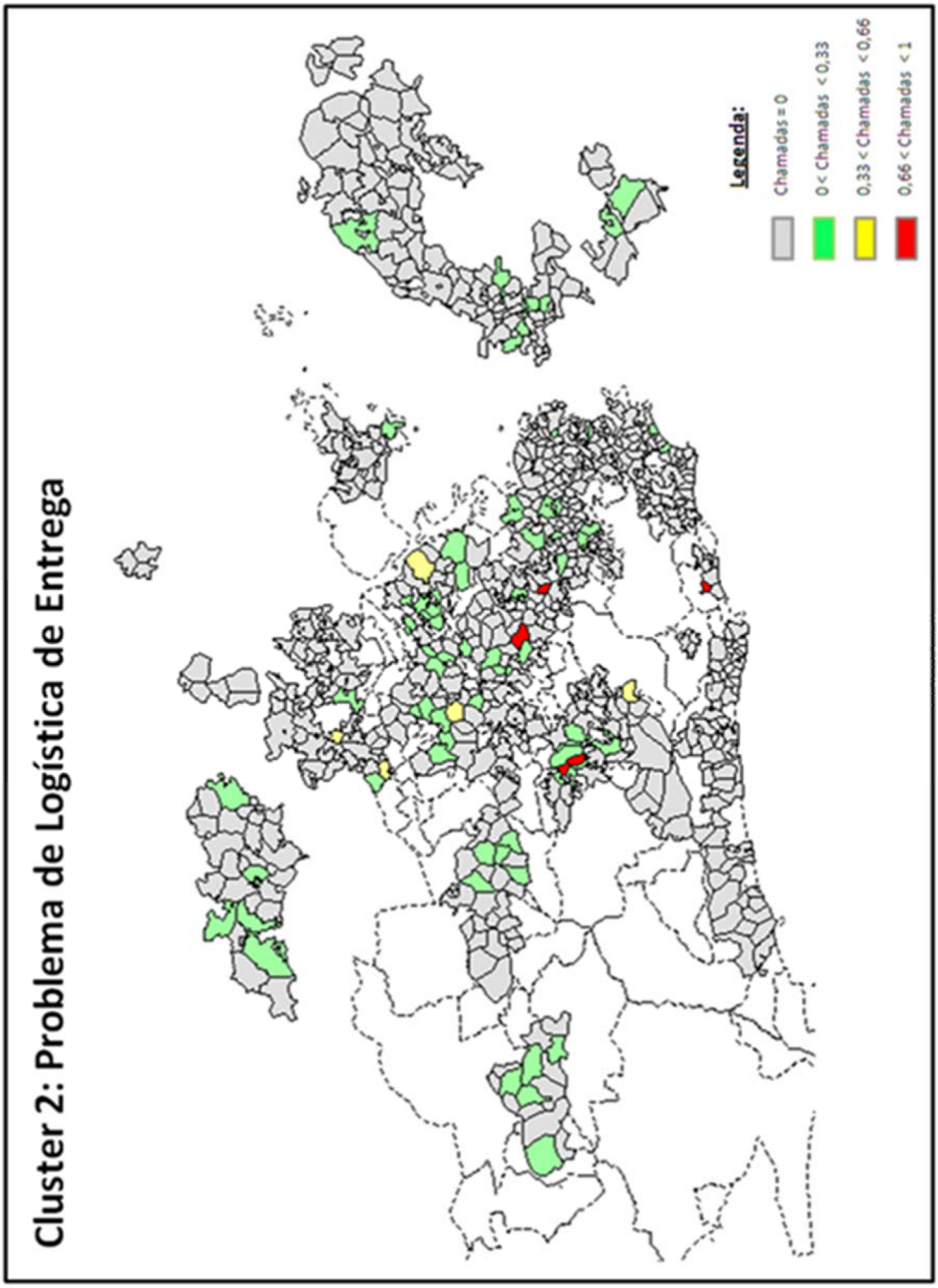

Cluster 2: Problema de Logística de Entrega
Legenda:
Chamadas = 0
0 < Chamadas < 0,33
0,33 < Chamadas < 0,66
0,66 < Chamadas < 1

Figura 48

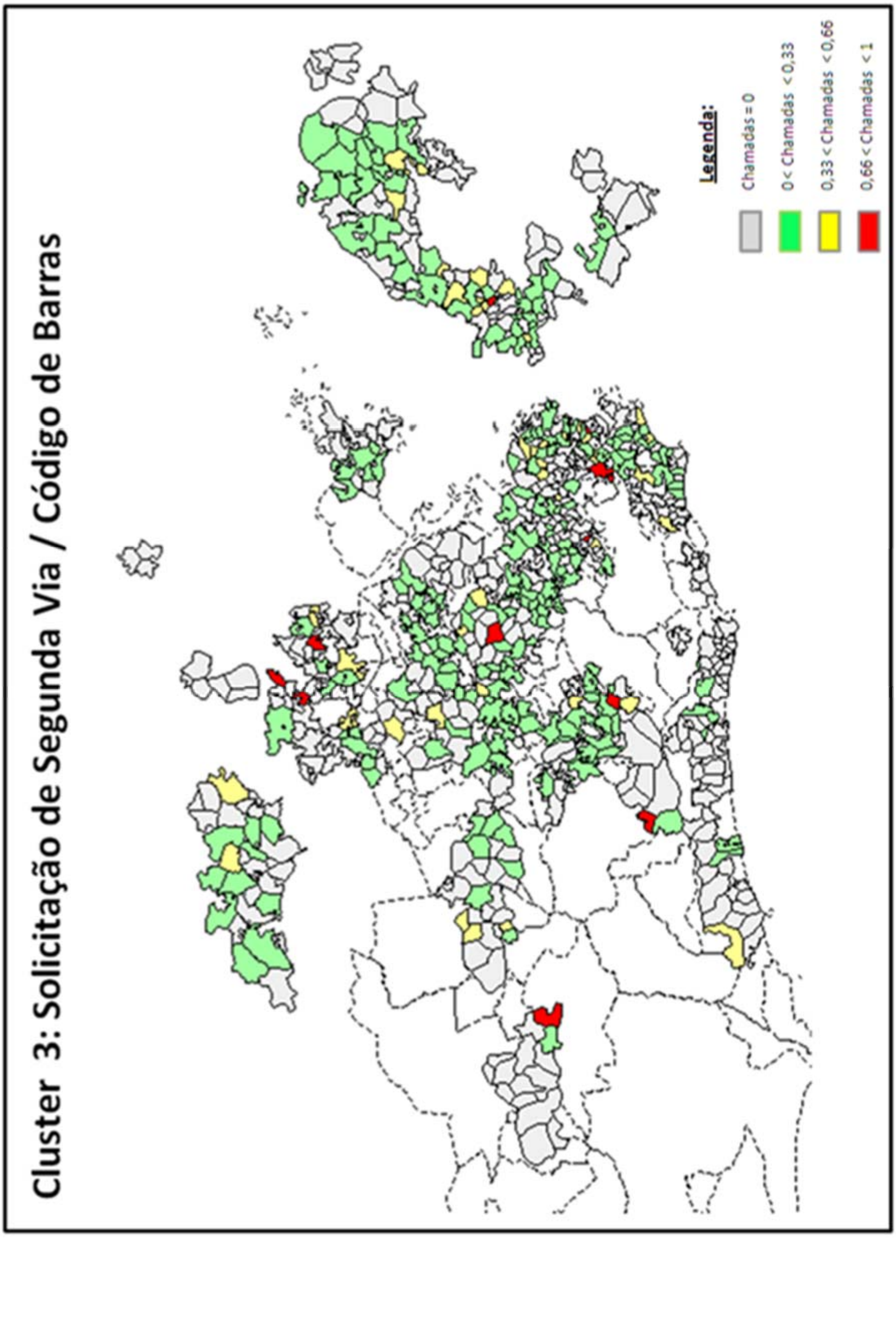

Cluster 3: Solicitação de Segunda Via / Código de Barras
Legenda:
Chamadas = 0
0 < Chamadas < 0,33
0,33 < Chamadas < 0,66
0,66 < Chamadas < 1

Figura 49

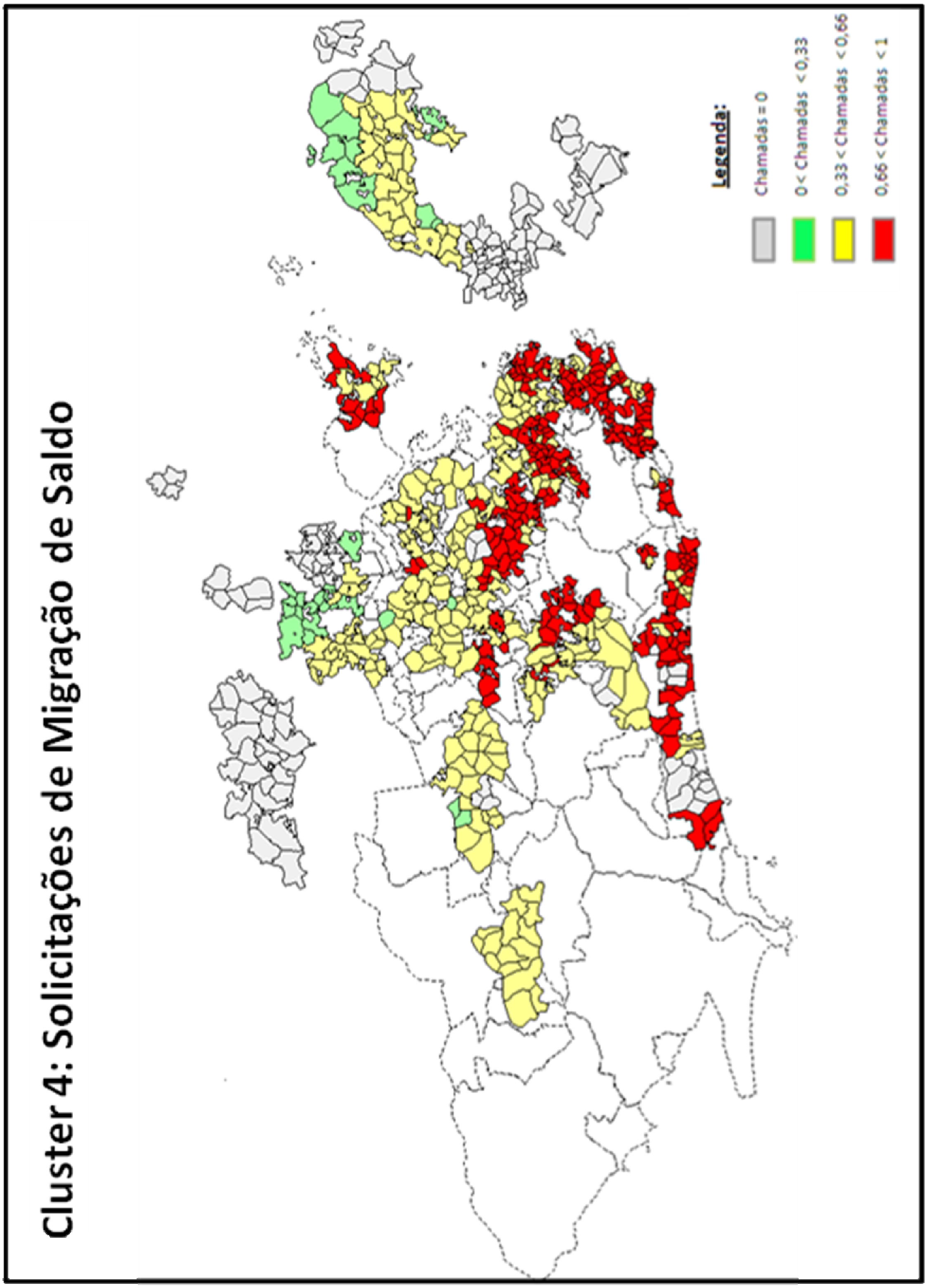

Cluster 4: Solicitações de Migração de Saldo
Legenda:
Chamadas = 0
0 < Chamadas < 0,33
0,33 < Chamadas < 0,66
0,66 < Chamadas < 1

Figura 50

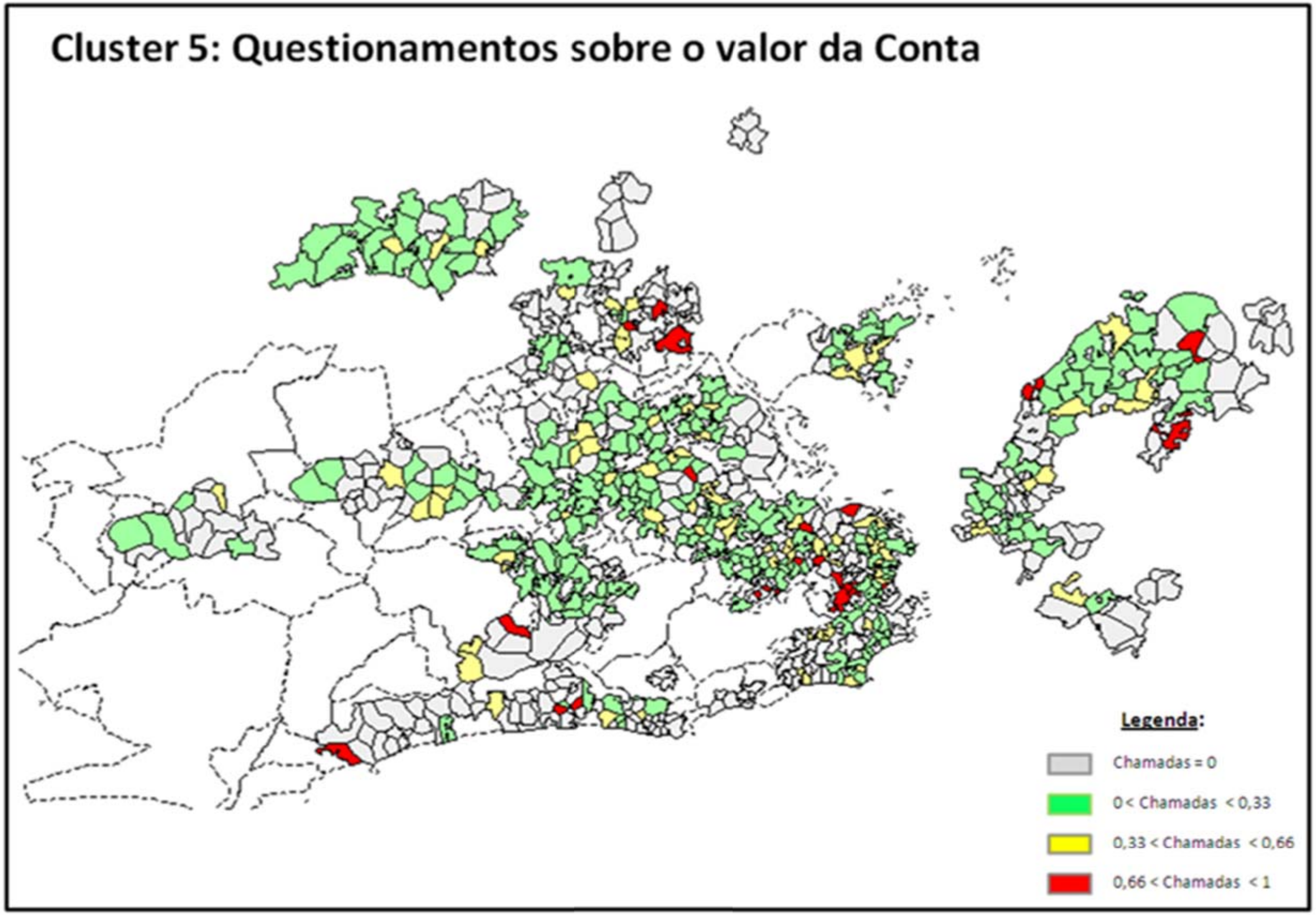

Cluster 5: Questionamentos sobre o valor da Conta
Legenda:
Chamadas = 0
0< Chamadas < 0,33
0,33 < Chamadas < 0,66
0,66 < Chamadas < 1

Figura 51

Com os Mapas Temáticos dos Clusters apresentados (Figuras 50 a 54) é possível tecer as seguintes conclusões:

Considerações sobre o Cluster 1 (Reparos)

Muito bem caracterizado na etapa de Text Mining, este *cluster* ao ter sua representação no Espaço Geográfico, permite identificar que os problemas de Reparo ,relatados pelos Clientes, não estavam, de maneira geral, limitados a uma região específica. Além disso, as taxas de incidência de problemas (relativizadas) também apresentaram-se baixas, dado que nesse Mapa Temático, em sua quase totalidade, apresentou taxas de reclamações classificadas como Reparos, em valores inferiores a uma relação de 0,33 reclamações por terminais instalados em um período totalizado de oito meses, conforme descrição preliminar do Banco de Dados Textual, utilizado neste Estudo de Caso.

Observa-se também que as incidências de Reparo, obviamente, não fazem distinção de Classe Social, pois se apresentam, de maneira geral, bastante uniformes - tanto nos Bairros da Zona Sul, Zona Norte, Centro, Niterói e Baixada Fluminense.

Exceção para o caso de um único ponto crítico onde a incidência de Textos Classificados como sendo de Reparos, foi superior à relação de 0,80 reclamações por terminal. Neste caso, com o auxílio do SIG é possível “religar” a camada de ruas, “desligada” até então para não poluir o entendimento mais genérico. Neste caso, como pode ser observado na Figura 52, esta área crítica (de Reparos) no ano de 2008 situou-se na região da Barra da Tijuca.

De forma a evidenciar o potencial do SIG como uma solução para tomada de decisão em empresas, observa-se na Figura 56, o detalhamento das ruas, com seus respectivos nomes evidenciados.

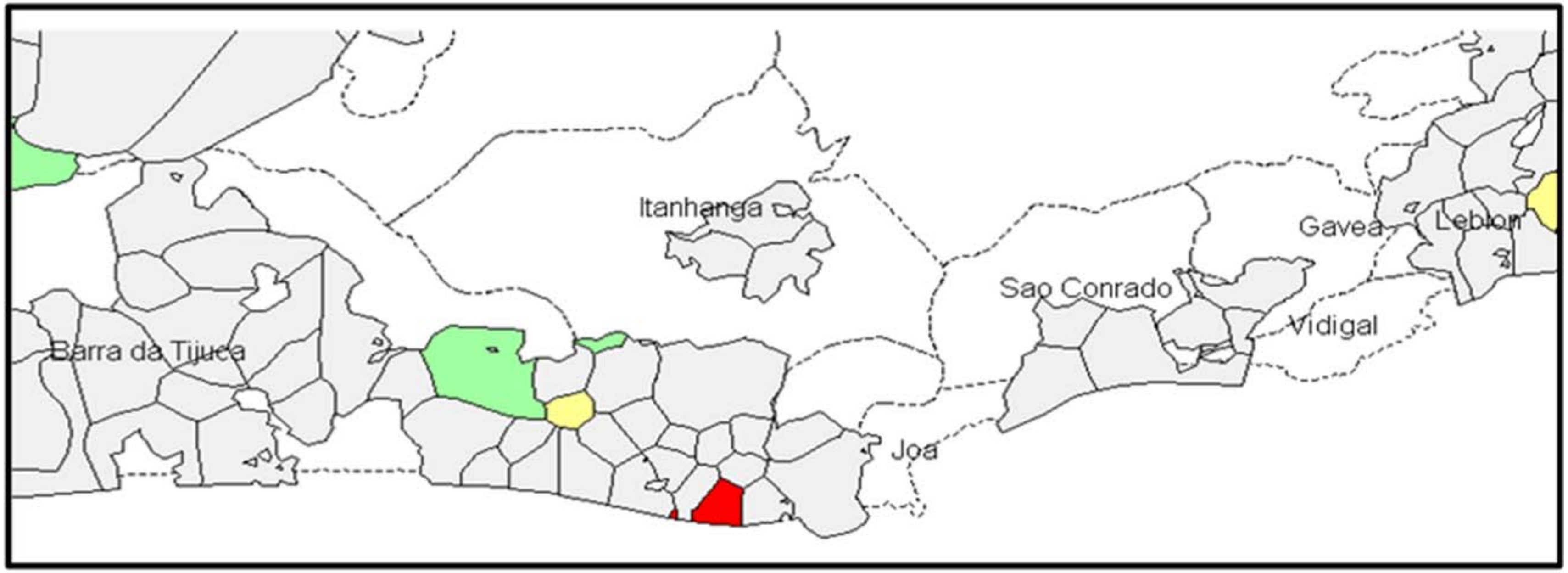


Figura 52: Detalhamento de Área Crítica do Cluster 1 (Reparos)

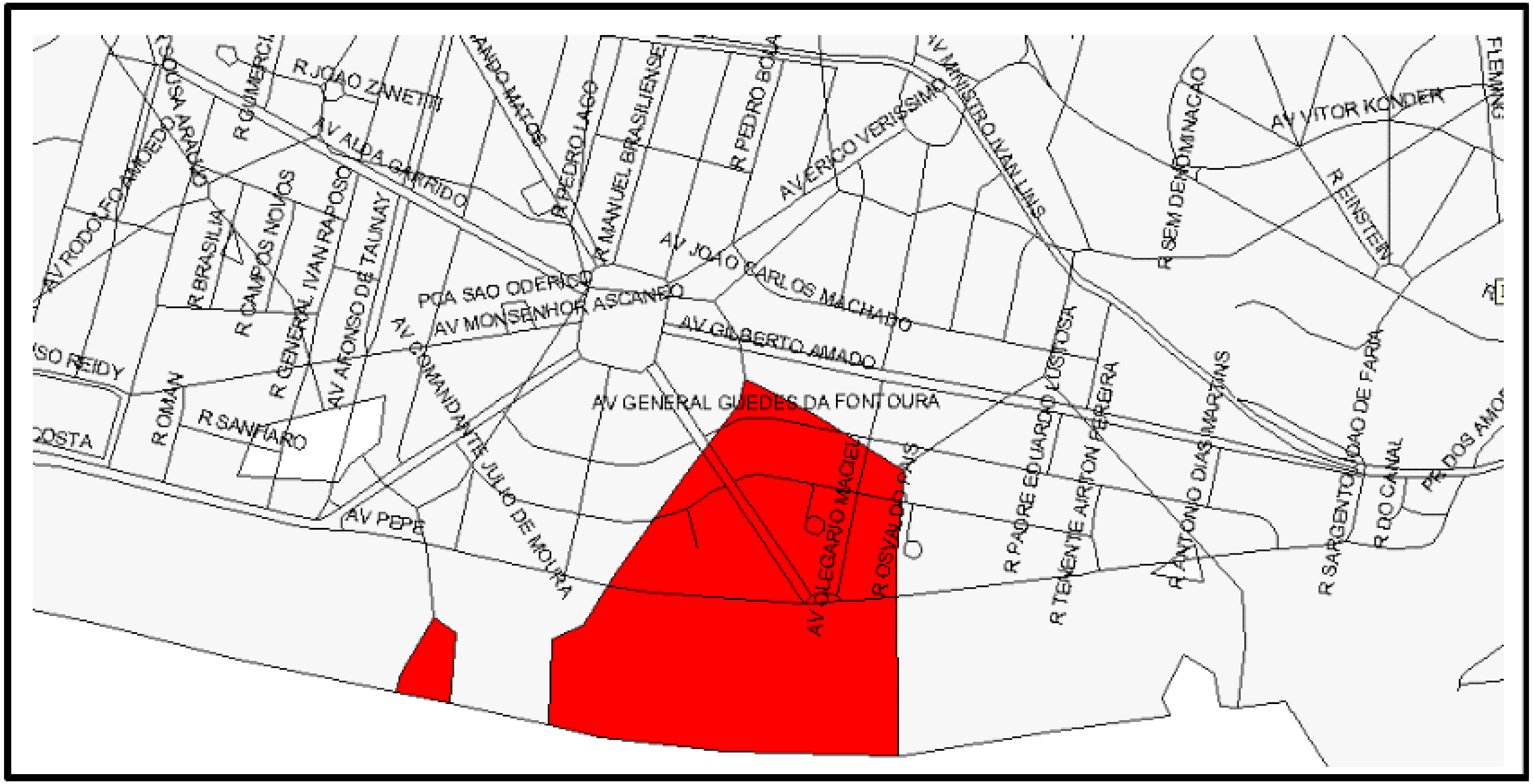


Figura 53: Detalhamento de Ruas Críticas do Cluster 1 (Reparos)

Considerações sobre o Cluster 2 (Problema de Logística de Entrega de Aparelho e Fatura de Clientes)

O que se pode observar a respeito da distribuição espacial das reclamações relacionadas a Problemas de Logística sejam elas referentes à entrega de aparelhos ou mesmo relacionados à entrega da Fatura Mensal (boleto para pagamento), é que as áreas onde ocorrem as maiores concentrações de reclamações são justamente áreas consideradas como áreas de risco, ou na proximidade de favelas - onde muitas vezes o trabalho dos Correios ou das Empresas de Logística fica demasiadamente moroso em função de problemas sociais e violência. Para exemplificar isto, observa-se a Figura 54.

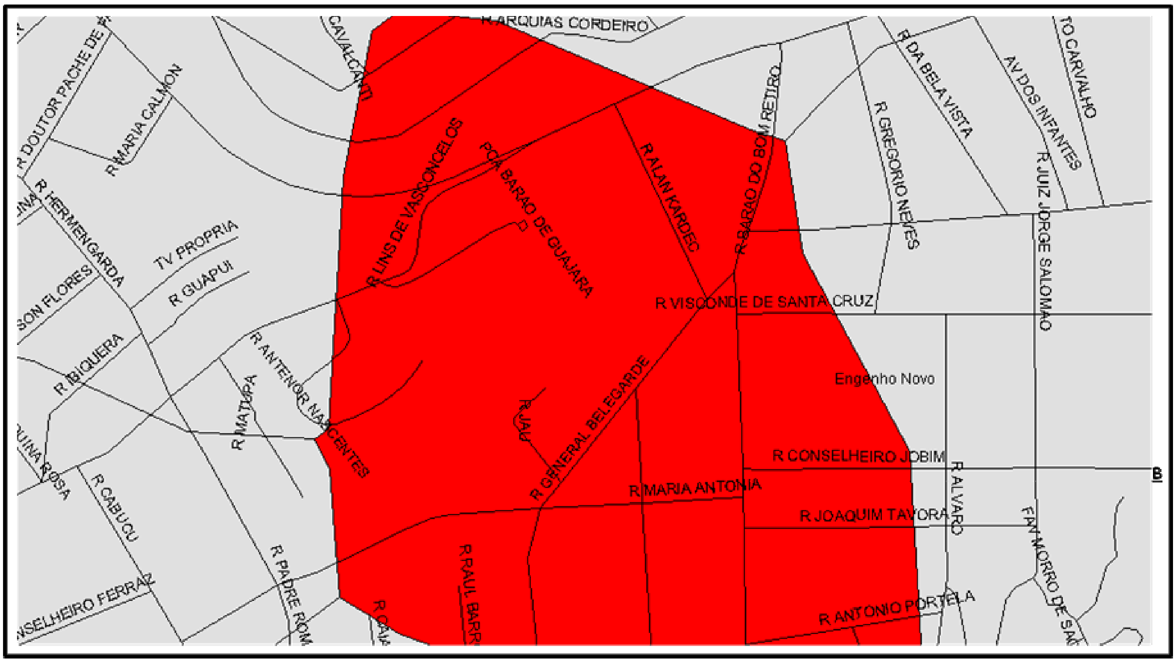


Figura 54: Detalhamento de uma das áreas críticas do Cluster 2

Ao observarmos uma foto de satélite da região evidenciada na Figura 58 (através do Google Maps) pode se verificar claramente que se trata de região com uma grande concentração de favelas, conforme premissas explicativas das áreas críticas deste Cluster.

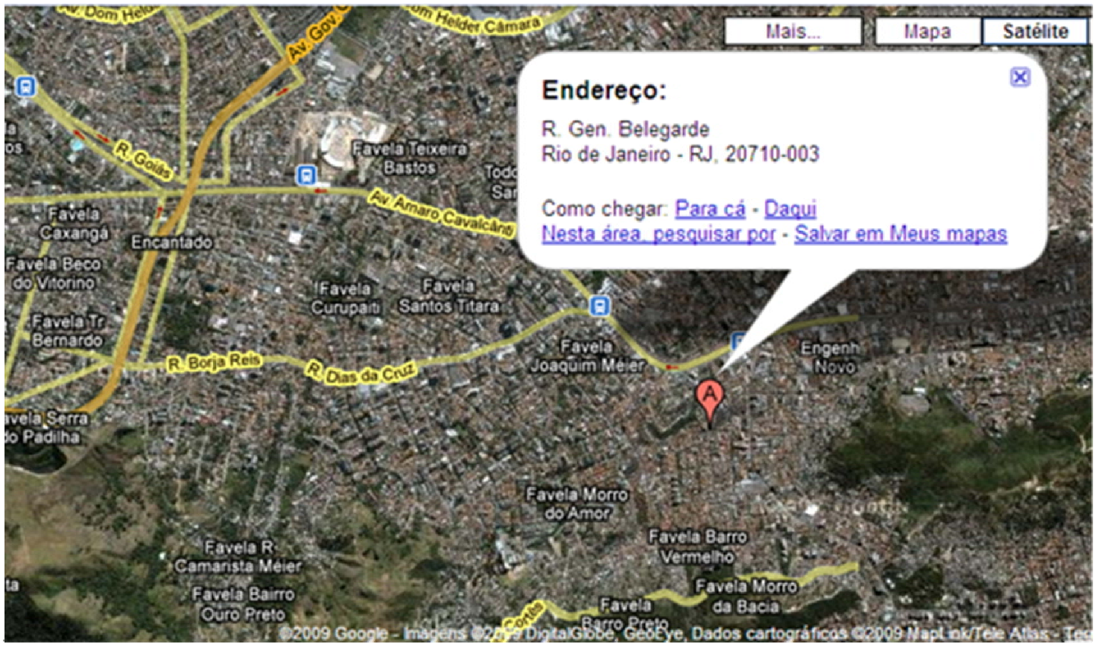


Figura 55: Foto de Satélite de Região Crítica do Cluster 2

Ainda exemplificando a aplicação da metodologia de georeferenciamento, com a utilização de Softwares com código aberto (Google Maps) observa-se na Figura 56 a "*plotagem*" de parte da área de Cobertura da Operadora que cedeu os dados - por questão de "confidencialidade" dos dados da Operadora, não foi possível mostrar neste material com um "zoom" maior. Entretanto, isto é possível de ser realizado com a aplicação de programação em KML, conforme descrito na revisão bibliográfica e metodologia. Nesta Figura (Figura 56), a área evidenciada com a cor verde representa a Cobertura e o fundo correspondente à uma foto de satélite. Desta forma fica plenamente comprovada a aplicabilidade da metodologia neste aspecto.

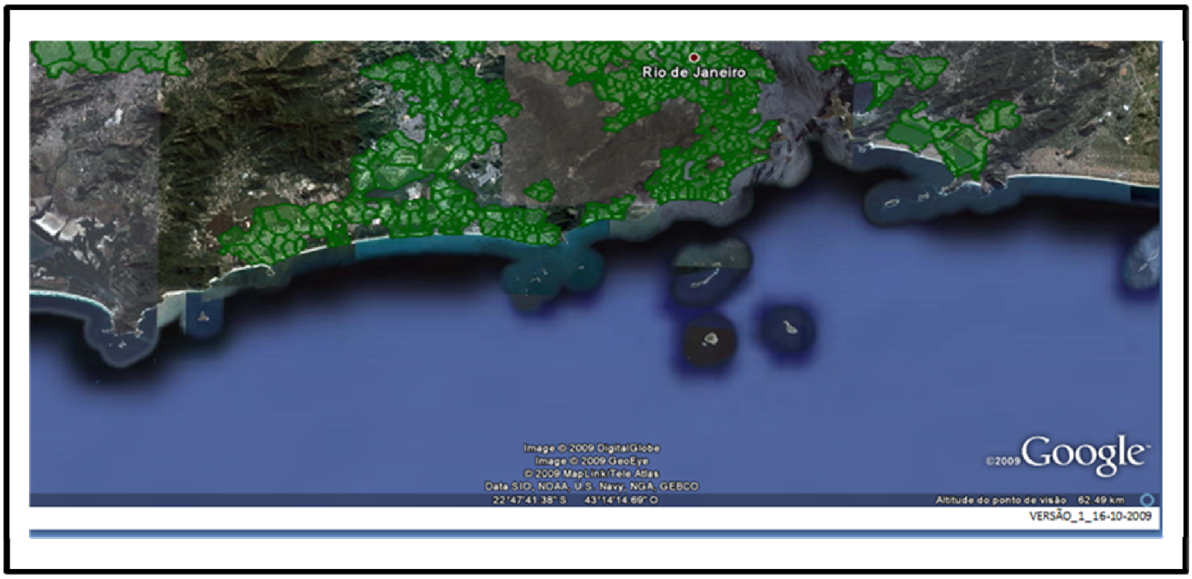


Figura 56: Visualização de parte da Cobertura através do Google Maps

<u>Considerações sobre o Cluster 3 (Solicitação de Segunda Via)</u>

A observação da dispersão de concentração das solicitações classificadas como "Solicitações de Segunda Via" remetem à hipótese de que este tipo de problema possa relacionado a algum tipo de problema sistêmico da Operadora, durante os meses em que os relatos de Clientes foram observados. Isto porque o mapa temático apresenta uma razoável uniformidade para a maior parte da Região Metropolitana do Rio de Janeiro e Baixada Fluminense. Contudo, pela experiência do Autor na área de Telecomunicações, este tipo de Solicitação normalmente também pode estar associado a problemas de ordem econômica que os consumidores possam vir a passar. Quando isto ocorre, tais clientes entram em contato com a Operadora para informar que não receberam a fatura e que por esse motivo precisam de um novo boleto. Normalmente, nestas situações, é estipulada uma nova data de vencimento e o consumidor "de certa forma" tem o seu problema de pagamento temporariamente resolvido. Neste aspecto (referente a possíveis problemas de ordem financeira dos clientes) cabe ressaltar que ao confrontar-se a área do mapa onde é maior a concentração de Classe AB (conforme Mapa Temático específico de Classe Social da Figura 34) verifica-se que justamente nesta área é pequena ou inexistente a concentração de solicitações de Segunda Via, corroborando a hipótese de que esta solicitação pode estar ligada a um artifício do consumidor para ganhar acréscimo de tempo para pagamento de suas contas. Nas Figuras 57 e 58 o detalhamento do que foi descrito acima.

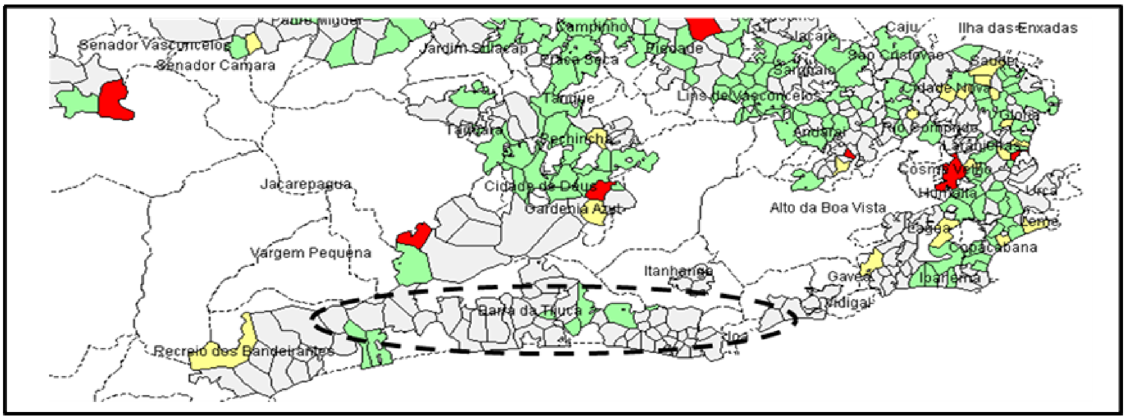


Figura 57: Detalhamento de área da Cidade com menor concentração de Seg. Vias

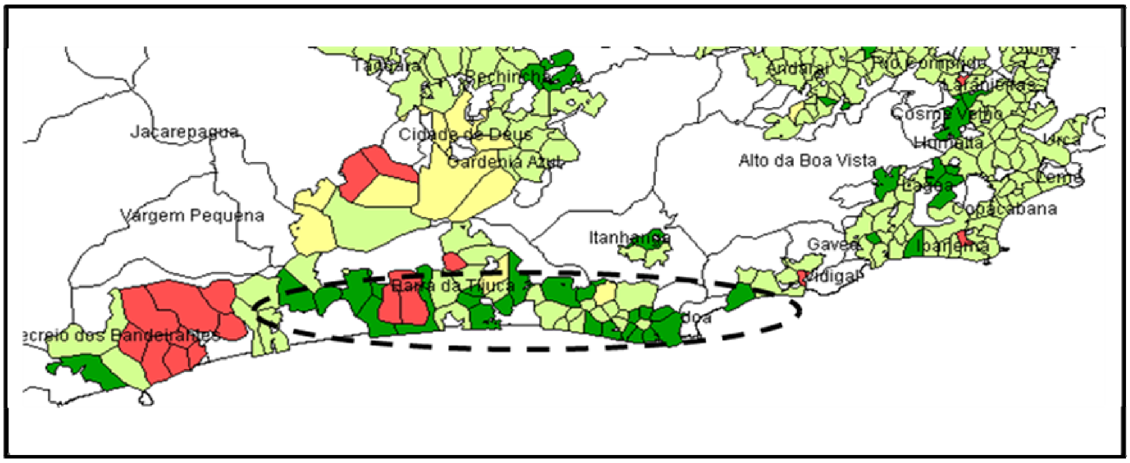


Figura 58: Detalhamento de área da Cidade com maior concentração de Classe AB

Considerações sobre o Cluster 4 (Solicitações de Migração de Saldo)

A observação do mapa temático do Cluster 4 evidencia uma forte diferenciação de ocorrências de problemas ou de solicitações de migração de saldo em áreas geográficas bem diferenciadas em relação às respectivas quantidades de reclamações. Contudo, praticamente todas as regiões tiveram consumidores ligando para o Call Center por este motivo, sobretudo as regiões da Zona Sul e Zona Norte, onde a quantidade de chamadas proporcionais à quantidade de clientes é muito maior. Em uma aplicação do recurso de Mineração de Textos combinada ao Georeferenciamento para uso comercial, ou seja, para una aplicação de melhoria de prestação de serviços em uma organização, este tipo de informação pode ser informação valiosa para melhoria de processos e de entendimento das necessidades de clientes.

Considerações sobre o Cluster 5 (Questionamentos sobre o Valor da Conta)

Neste cluster as reclamações também aparecem razoavelmente bem distribuídas, contudo, um fenômeno interessante pode ser observado. Algumas das áreas onde a ocorrência foi mais crítica, ou seja, as áreas demarcadas no mapa temático com as cores amarela e vermelho, correspondiam na época a ERBs que haviam sido implantadas em um período de até um ano, ou seja, eram Estações Radio Base relativamente novas. Esta informação pode indicar que os problemas sobre questionamentos de valor de conta ocorreram de maneira uniforme nas áreas demarcadas no mapa, mas a ocorrências de maior intensidade foram em regiões com novos clientes (das novas ERB's) que ainda não estavam habituados à sistemática de cobrança pelos serviços prestados pela Operadora de Telecomunicações.

5.5 RESUMO DO CAPÍTULO

O Estudo de Caso evidenciou de uma forma prática e objetiva toda a metodologia proposta, tanto na etapa de Mineração de Textos quanto na etapa de Georeferenciamento. Além disso, com a aplicação da metodologia em um problema real o (o processamento a partir dos registros das chamadas de Clientes), ficou visível o valor que pode ser agregado aos negócios de organizações que optem por utilizar tais tecnologias como ferramentas de suporte a tomada de decisões.

As principais dificuldades encontradas foram relativas ao Pré-Processamento das bases Textuais disponibilizadas: de difícil manipulação, com muitos erros de digitação, com uso de expressões não convencionais e de caráter técnico (que só fazem sentido para a realidade empresarial em questão), além do eventual "*viés*" do Operador de Call Center que evita, de certa maneira, o uso de expressões que denotem insatisfações (no caso de transcrição de reclamações por Operadores de Call Center). Resolvido este grande desafio, de transformar a base de dados "extremamente desestruturada" em algo passível de processamento - através da aplicação de diversas técnicas respaldadas na pesquisa bibliográfica, foi possível evidenciar um processo de Mineração de Textos do nício ao fim, permitindo ao final deste, a aplicação de técnicas de *clusterização*, que evidenciaram diferenciação entre grupos de clientes.

Na etapa de Georeferenciamento, a aplicabilidade da proposta também ficou evidenciada através da utilização de várias técnicas descritas na revisão bibbliográfica e metodologia. Contudo, apenas uma possibilidade não foi explorada, apesar de ter sido referenciada na revisão bibliográfica: o uso de técnicas de Interpolação (Krigagem, por exemplo) para a criação de camadas referentes à características obtidas do conjunto de textos. Em função do prazo para execução deste trabalho isto não foi realizado – mas em futuros trabalhos pode ser uma alternativa interessante, caso não exista uma delimitação de áreas geográficas bem definida e com características semelhantes, como é o caso das Estações Radio Base - menor granularidade geográfica utilizada no trabalho.

6. CONCLUSÕES E SUGESTÕES PARA FUTURAS PESQUISAS

A aplicação da metodologia proposta, com suas particularidades de integração de conhecimentos e de ferramentas, permitiu vislumbrar um enorme potencial de futuras pesquisas e aplicações nestas duas áreas do conhecimento combinadas (Mineração de Textos e Georeferenciamento), permitindo assim um ganho significativo de conhecimento sobre a opinião de consumidores; sobre o que falam ou escrevem e a influência do espaço geográfico e dos grupos sociais.

As principais dificuldades encontradas no desenvolvimento deste trabalho foram as relacionadas à manipulação das bases textuais obtidas para o Estudo de Caso. Estas bases, que eram muito extensas – cerca de 15 milhões de registros – continham muitos caracteres estranhos, muitas abreviações, termos técnicos e uma "*pseudo-linguagem*" característica dos Operadores de Call Center. Por tudo isso, as etapas de Limpeza e de Pré-Processamento foram particularmente complexas e demoradas. Futuros trabalhos que venham a ser realizados a partir da leitura desta Tese poderão se beneficiar com a utilização do "workflow" sugerido na mesma.

Algo que também ficou nítido no Estudo de Caso e que também precisa ser considerado em futuras análises similares é o fato de o Operador de Call Center, que pela natureza de seu trabalho, é constantemente monitorado, não ter total "liberdade" para escrever na íntegra os comentários e reclamações dos consumidores que entram em contato com o Call Center. Este fato pode levar a eventuais "*falsas conclusões*". Portanto, o pesquisador deve estar atento às nuances dos textos e das palavras-chave de reclamações que porventura possam ter sido "atenuadas" de maneira intencional.

Além disso, a partir deste trabalho é possível elaborar outros questionamentos relevantes: a maneira como os consumidores interagem com empresas e instituições, através de interações por escrito (textos) está correlacionada ao seu nível sócio-econômico-cultural? Esta Tese mostrou que existem correlações perceptíveis (evidenciadas no Estudo de Caso), mas que precisam ser investigadas mais profundamente

Observa-se que os "atores"; redes sociais, indivíduos e instituições, interagem uns com os outros, todo o tempo. O desafio da ciência é conseguir mensurar, identificar, isolar, todo este processo (de interação) com respaldo dos pesquisadores da área de Ciências Humanas, e agora também com o respaldo de pesquisadores da área de Data Mining e de Georeferenciamento, de forma a melhor entender a intensidade de tais "*influenciações*". Este parece ser apenas o começo de descoberta de inter-relações em áreas fascinantes e multidisciplinares do conhecimento, que estão cada vez mais em evidência nos dias de hoje.

Como sugestão de futuras pesquisas, pode ser interessante o aprofundamento no uso de códigos abertos nos SIG's. Durante a etapa de pesquisa foi possível observar a crescente disponibilização de conteúdo pelos usuários da Web sobre locais georeferenciados, com suas particularidades, como por exemplo, a localização de escolas, empresas, hospitais, "*shopping centers*", etc. Ou seja, já existe informação de valor agregado de todos estes locais e suas respectivas características sócio-econômicas, que podem ser objeto de análise de Mineração de Dados - de forma a possibilitar sobre possíveis oportunidades de negócios ou de melhorias sociais para a população residente (nestas áreas).

Nesta Tese não foram utilizados os recursos de interpolação, explicado em detalhes na revisão bibliográfica (*Krigagem*) em função da não existência de tal recurso na versão de software (SIG) utilizada. Contudo, para os mapa temáticos elaborados neste Estudo de Caso foi possível obter uma granularidade razoável de visualização em função de as análises terem sido realizadas a partir das Estação Radio Base, que por sua natureza, comportam um número limitado de consumidores.

Algo que também pode ser relevante: incorporar as coordenadas (x, y) como variáveis para a etapa de clusterização, ou seja, utilizando conjuntamente as variáveis obtidas no processo de Mineração de Textos com as variáveis georeferenciadas.

Finalizando, é possível afirmar que, a partir de trabalho, com a utilização dos processos detalhadamente descritos, muitas possibilidades futuras de descoberta de conhecimento relevante, poderão agregar valor ao conhecimento aprofundado dos consumidores por parte de empresas e instituições.

REFERÊNCIAS BIBLIOGRÁFICAS

AMARAL S.A. (1998).Marketing: abordagem em unidades de informação. Brasília: Thesaurus.

APPELT, D. E; Israel, D. J. (1999). Introduction to Information Extraction Technology. In Proceedings of the 16th International Joint Conference on Artificial Intelligence.

BILMES J. A. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. International Computer Science Institute. Berkeley.

BERRY, M. W.(2004). Survey of text mining: clustering, classification, and retrieval. Springer-Verlag : New York.

BOND, C. S. (2004). Web Users Information Retrieval Methods and Skills. Emerald Online Information Review. Vol. 28 - Number 4 – pp 254 /259

BORKO, H., BERNICK M. (1963). Automatic document classification. In Journal of the Association for Computing Machinery, 10:151—162..

BURROUGH, P. A.(1986) Principles of geographical information Systens for Land Resources Assessment. New York: Oxford University Press.

CÂMARA G., PEDROSA B.M.. (2004) Modelagem Dinâmica e Geoprocessamento. INPE, 8561-PRE/4305.

CAMARGO, E.C.G. (1997). Desenvolvimento, implementação e teste de procedimentos geoestatísticos (krigeagem) no sistema de processamento de informações georreferenciadas (Spring). São José dos Campos.. Dissertação (Mestrado) - Instituto Nacional de Pesquisas Espaciais.

CAMARGO, Y. B. L. (2007) Abordagem Linguística na Classificação de Textos em Português. Dissertação de Mestrado – Engenharia Elétrica – COPPE/UFRJ.

CHANG H., CHUNG H. L. (2005). A Mineração de TextosApproach for Automatic Construction of Hypertexts. Chang Jung University, Taiwan e National Kaoshiung University of Applied Scieneces, Taiwan.

CARVALHO, J.R.P., SILVEIRA, P.M. da; VIEIRA, S.R.(2002) Geoestatística na determinação da variabilidade espacial de características químicas do solo sob diferentes preparos.

CRISTIANINI, N. & SHAWE-TAYLOR, J. (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.Cambridge, GB. 2000.

CZINKOTA, Michael R. Et al. (2001) Marketing: as melhores práticas. Porto Alegre: Bookman, p.278-279

CHIDANAND, A.; DAMERAUF, Weiss, S.M. (1994) Automated Learning of Decision Rules for Text Categorization - ACM Transaction on Information Systems, Vol.12, No.3.

CHIEN L.F., LEE H.J., LU W.H. (2004). Anchor Mineração de Textosfor Translation of Web Queries: A Transitive Translation Approach. Academia Sinica e National Chiao Tung University.

COUCLELIS H. (1997). From Cellular Automata to Urban Models: New Principles for Model Development and Implementation

DEERWESTER et al. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, v.41, 1990, pp. 391-407.

DEMPSTER, A., LAIRD, N., and RUBIN, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1– 38

DOLFUS, O. (1991). O Espaço Geográfico. Rio de Janeiro, Bertrand Brasil.

DIRKS, K.N.; HAY, J.E.; STOW, C.D. & HARRIS, D. (1998) Highresolution studies of rainfall on Norfolk Island Part II: Interpolation of rainfall data. J. Hydrol., 208:187-193.

DUMAIS S. (1998). Inductive Learning Algorithms and Representations for Text Categorization. Microsoft Research Center – Redmond / USA.

EGENHOFER, Max J., FRANZOSA, Robert D. (1991) *Point-set topological spatial relations*. International Journal of Geographical Information Systems, London, v.5, n.2, p.161-174.

FIGUEIREDO L. H., STOLFI J. (2004). Lecture Notes on the EM Algorithm. Technical report, Institute of Tele-communication.

FIGUEIREDO M. A. Z. (2008). Mineração de Dados Não Estruturados e Temporais. Dissertação de Mestrado. UFRJ/COPPE.

FELDMAN, R; Dagan I. (1995). Knowledge discovery in textual databases (KDT) - in proceedings of The First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal Canada, August 20-21, AAAI Press, 112-117.

GALHO T. S., MORAES S.M.W. (2003). Categorização Automática de Documentos de Texto Utilizando Lógica Difusa. Universidade Luterana do Brasil.

GERSHTEYN Y. (2005). Use of SPEDIS Function in Finding Specific Values. On line em http://www2.sas.com/proceedings/sugi25/25/cc/25p086.pdf

GOMES, J.M., VELHO, L. (1995) Computação Visual: Imagens. Rio de Janeiro, SBM.

GOTWAY, C.A.; HARTFORD, A.H. (1996). Geostatistical methods for incorporating auxiliary information in the prediction of spatial variables. J. Agric., Biol. Environ. Statis., 1:17-39.

HAN J., KAMBER M. (2006). Data Mining, Second Edition, Second Edition : Concepts and Techniques The Morgan Kaufmann Series in Data Management Systems

HARTIGAN J.A.(1972). Direct Clustering of a Data Matrix. Journal of the American Statistical Association.

HU M., LIU B.(2004). Mining Opinion Features in Customers Reviews.

HUIJBREGTS, C.J.(1975). Regionalized variables and quantitative analysis of spatial data. In: Davis, J.C. & McCullagh, M.J. (ed) Display and anaysis of spatial data. New York, John Wiley.

ISAAKS, E.H.; SRIVASTAVA, R.M (1989) An Introduction to Applied Geostatistics. New York: Oxford University Press.

JOURNEL, A.G.; HUIJBREGTS, Ch.J. (1978) Mining Geostatistics. London: Academic Press.

KOHONEN, T. (2001). Self –Organizing Maps. Terceira Edição. New York: Springer-Verlag.

KONCHADY M. (2006).Mineração de TextosApplication Programming.

KOPPEl, M., & SHCLER, J. (2005). The importance of Neutral Examples for Learning Sentiment. In Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations (FINEXIN).

________________________ (2005). Using Neutral Examples for Learning Polarity. In Proceedings of IJCAI.

KOSALA and BLOCKEEL (2000). Web mining research: a survey, SIGKDD Explorations: Newsletter of the Special Interest Group SIG on Knowledge Discovery & Data Mining, 2.

KU, L.-W., LEE, L.-Y., WU, T.-H., & CHEN, H.-H. (2005). Major Topic Detection and its Application to Opinion Summarization. In SIGIR 2005 (pp. 627–628).

LANBIN, E.F. (1994). Modeling Deforestation Proceses – A Review, Trees series B: Research Report. European Commission, Louxembourg

LAUFER P.L.(2003), Sistemas Fuzzy com Matlab. Departamento de Engenharia Eletrônica e de Computação – Escola Politécnica – UFRJ.

LI T. et al (2004).Document Clustering via Adaptative Subspace Iteration.Annual ACM Conference on Research and Development in Information Retrieval - 27th annual international ACM SIGIR conference on Research and development in information retrieval table of contents. Session: Clustering table of contents - Pages: 218 - 225

LIPPMANN, Richard (1987). An Introduction to Computing with Neural Works. IEEE Computer Society, v. 3, n. 4, p. 4-22.

LIU B. (2006), Web Data Mining: Exploring Hyperlinks, Contents and Usage Data Ed. Springer Verlag, NY.

LIU, B., HU, M., & CHENG, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th international World Wide Web Conference (WWW-2005) (pp. 10–14). Chiba, Japan: ACM Press.

LOVINS, J. B. (1968) Development of a stemming algorithm. Mechanical Translation and Computacional Linguistics, Volume 11, Number 1-2, pages 22-31.

LUNA, J. E. O. (2004). Algoritmos de Expectation Maximization para Aprendizagem de Redes Bayesianas a partir de Dados Incompletos. Dissertação de Mestrado- Universidade Federal do Mato Grosso do Sul.

MACHADO, R.V. (1994). Variabilidade espacial de atributos físico-hídricos em uma hidrosequência de solos bem à muito mal drenados. Dissertação (Mestrado), Escola Superior de Agricultura de Lavras.

MANNING C., SCHUTZE, H. (1999). Foundations of Natural Language Processing. MIT Press.

MATHERON J. (1965) Les Variables Régionalisées et Leur Estimation: une Application de la Théorie des Fonctions Aléatoires aux Sciences de la Nature, Tese de Doutorado, Masson, Paris.

McCALLUM A.K., NIGAM K.(1998). A Comparison of Event Models for Naïve Bayes Text Classification. In: Proceedings of the First AAAI Workshop on Learning for Text Categorization, pages 41-48, Madison, USA.

MAYERA D., LEISCH F., HORNIB K. (2003) The Support Vector machine under Test

A Institut f_ur Statistik und Wahrscheinlichkeitstheorie, Technische Universit_at Wien, Wiedner Hauptstra_e 8-10/1071, A-1040 Wien, Austria - Institut f_ur Statistik, Wirtschaftsuniversit_at Wien, Augasse 2-6, A-1090 Wien, Austria, online em http://www.sciencedirect.com

MELLO, C. R.; LIMA, J.M.; SILVA, A. M.; MELLO, J. M.; OLIVEIRA, M. S. (2003) Krigagem e Inverso do Quadrado da Distancia para Interpolação dos Parâmentros da Equação de Chuvas Intensas. R. Bras. Ci. Solo. V. 27, p. 925-933.

MIRKIN B.(1996).Mathematical Classification and Clustering. Kluwer Academic Publishers. ISBN 0792341597

MONTEIRO A.M.V. & CAMARA G.(2007). Conceitos Básicos em Ciência da GeoInformação. Online na página do Inpe: http://www.dpi.inpe.br/gilberto/introd/cap2-conceitos.pdf

MOREIRA, O. V; Huyck C.R.(2001). Stemming Algorithm for the Portuguese Language. In Proceedings of the SPIRE conference, Laguna de San Raphael, Chile.

MURTAGH, F. (1983). A survey of recent advances in hierarchical clustering algorithms.The Computer Journal 26 (40)

NASUKAWA T., NAGANO, T.T.(2001) "Knowledge discovery using robust natural language processing", Pacific Association for Computational Linguistics (PACLING), pp.189-198.

NITIN I., ZHAN, GT., DAMERAU F., WEISS S.(2005). Text Mining: Predictive Methods for Analyzing Unstructured Information.

OCHI L. S., DIAS C.R., SOARES S.S.(2004). Clusterização em Mineração de Dados. Intituto de Computação – Universidade Federal Fluminense.

PAPADIAS, Dimitris, THEODORIDIS, Yannis (1997) Spatial Relations, Minimum Bouding Rectangles, and Spatial Data Structures. International Journal of Geographical Information Systems, London, v.11, n.2, p.111-138.

PANG L. et al (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. EMNLP.

PORTER, M. (1980). An algorithm for suffixing stripping. Program, Volume 14, Number 3, pages 130-137.

SALTON, G., WANG, A., & YANG, C. S. (1975). A Vector Space Model for Information Rretrieval. Journal of the American Society for Information Science, 18(11), 613–620.

SALVIANO, A.A.C. (1996).Variabilidade de atributos de solo e de Crotalaria juncea em solo degradado do município de Piracicaba-SP. Piracicaba, Tese (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

SANTORINI B.(1990). Part of Speech Tagging Guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

SILVA, A.P.(1988) Variabilidade espacial de atributos físicos do solo. Piracicaba, 105p. (Doutorado - Escola Superior de Agricultura "Luiz de Queiroz/USP).

SILVA C. F., VIEIRA R., OSÓRIO F.S. (2003). Uso de Informações Linguístiscas em Categorização de Textos utilizando Redes Neurais Artificiais. PIPCA – UNISINOS / RS.

SAN JUAN E., FIDELIA Ibkwe (2006). Mineração de TextosWithout Document Context. Université de Metz e Université de Lyon.

SCHIESSL, J. M. (2007). Descoberta de Conhecimento em Textos aplicada a um Sistema de Atendimento ao Consumidor. Dissertação de Mestrado, UNB.

SEBASTIANI, F., 1999, "A Tutorial on Automated Text Categorization". In: Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, pp.7-35, BuenosAires, AR.

SEMOLINI, R. (2002). Support Vector Machine, Inferência Trandutiva e o Problema de Classificação. Dissertação de Mestrado, UNICAMP.

SOUZA, L.C. (1999). Variabilidade espacial da salinidade de um solo aluvial no semi-árido paraibano. Campina Grande. Dissertação (Mestrado) - Universidade Federal da Paraíba

SOUZA, L.S.(1992).Variabilidade espacial do solo em sistemas de manejo.Tese (Doutorado) - Universidade Federal do Rio Grande do Sul.

SPERANDIO M., DUTRA R., COELHO J.(2009). O Método Ward de Agrupamento de Dados e sua Aplicação com os Mapas Auto-Organizáveis de Kihonen. On Line em www.labplan.ufsc.br/sperandio em 13/06/2009.

SULLIVAN, D. (2001). Document Warehousing and Text Mining. John Wiley & Sons, New York.

TAN, Ah-Hwee (1999). Text mining: the state of the art and the challenges. In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases-PAKDD'99, Beijing, 65-70.

TANG H., TAN S., CHENG Y. (2009) A Survey on Sentiment Detection of Reviews. Expert Systems with Applications.On Line em www.elsevier.com/locate/ewsa.

THALES S. K., FONSECA L. M. G., BAÇÃO F. L. (2008) Expectation-Maximization x Self-Organizing Maps for Image classification. Signal Image Technology and Internet Based Systems, 2008. SITIS '08. IEEE International Conference on. *Bali, Indonesia.*

THEODORIDIS, S. and KOUTROMBAS, K. (2003). Pattern Recognition. Academic Press.

TERVEEN L., et al. (1997). A System for Sharing Recommendations. Communications of the ACM, 40(3), 59–62.

TURNEY P.(2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews – Meeting of the Association for Computational Linguistiscs (ACL'02) – pags. 417-424.

VASILEIOS, H.; Gravano, L.; Maganti, A.(2000). "An Investigation of Linguistic Features and Clusters Algorithms for Topical Document Clustering". Proceedings of the

23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

VIEIRA, S. R. (2000). Geoestatística em estudos de variabilidade espacial do solo
In: NOVAIS, R. F. de; ALVAREZ V., V. H.; SCHAEFER, C. E. G. R. (Ed.). Tópicos
Em Ciência do solo. Viçosa: Sociedade Brasileira de Ciência do Solo, 2000. v. 1, p.
1-54

WARD J. H. (1963). Hierarchical Grouping to Optimiz an Objective Function. Journal of American Statistical Association, 58(301), 236-244

WITTEN (2004). Adaptative text mining: inferring structure from sequences". Journal of Discrete Algorithms 2, pp. 137-159, 2004.

YANG H. C., LEE C.H. (2004). A ext Mining Approach on Automatic Generation of Web Directories and Hierarchies. Chang Jung University, Taiwan e National Kaoshing University of Applied Sciences, Taiwan.

YANG, Y.; PEDERSEN, J.P.(1997). A Comparative Study on Feature Selection in Text Categorization. Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp. 412-420.

WIEBE, J. (1994). Tracking point of view in narrative. Computational Linguistics, 233–287.

ZHAI C., LAFFERTY J. (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval – Carneggie Mellon University.

APÊNDICE A:

Macro em Excel para transformação do arquivo no formato “1 Texto = 1 Linha”

```
 Sub Salvando_Linhas()

' Macro gravada em 22/1/2009

Dim reg As String
Dim linha, linha2, i As Integer
Dim equipe(300) As String
Dim data(300) As String

Application.ScreenUpdating = False
Application.StatusBar = "Criando Planilhas"

linha = 2
linha2 = 2

While ActiveSheet.Range("i" & linha2) <> ""

    reg = ActiveSheet.Range("i" & linha2).Value
i = 0

            MyPath = "C:\Arquivo_XLS"
            Myname = "Reg " & reg
            Workbooks.Add
            Application.CutCopyMode = False
            ActiveWorkbook.SaveAs FileName:= _
            MyPath & "\" & Myname, FileFormat:=xlNormal, _
            Password:="", WriteResPassword:="", ReadOnlyRecommended:=False, _
            CreateBackup:=False
            Application.DisplayFormulaBar = True
            Windows("Matriz Reg.xls").Activate

While ActiveSheet.Range("b" & linha).Value = reg

        If ActiveSheet.Range("b" & linha).Value <> "" Then

            i = i + 1
            data(i) = ActiveSheet.Range("c" & linha)
        Windows("Reg " & reg & ".xls").Activate
        'Cells(i, 1) = equipe(i)
        'Cells(i, 2).Select
        Cells(i, 1) = data(i)
        'Selection.NumberFormat = "General"
        Cells(i, 2) = reg
        Windows("Matriz Reg.xls").Activate
        End If

    linha = linha + 1

    Wend

Windows("Reg " & reg & ".xls").Activate
ActiveWorkbook.Save
ActiveWorkbook.Close
linha2 = linha2 + 1
Wend
MsgBox "Fim"
End Sub
```

APÊNDICE B:

Sintaxe para conversão de arquivos no formato ".SAS7BDAT"

```
%tmfilter (dataset=mylib.tmoutds,
    dir=\\aclientmachine\Public\TM\Examples,
    destdir=\\aclientmachine\Public\TM\mydatasets,
    ext=doc txt html pdf,
    language= Portuguese english spanish,
    numbytes=20);
```

Onde:

dir: especifica o diretório onde os arquivos serão processados
destdir: especifica o diretórios onde o resultado do processamento será armazenado
ext: especifica o tipo de arquivo que poderá ser processado
language: especifica os idiomas que podem ser reconhecidos

APÊNDICE C:

Pesos das Variáveis SVD nos Clusters Finais

Cluster	Percentage	_SVD_1	_SVD_2	_SVD_3	_SVD_4	_SVD_5	_SVD_6	_SVD_7	_SVD_8
1	29%	0.28582	-0.08622	0.02931	0.12224	0.001238	0.09762	0.02244	0.01087
2	43%	0.24522	-0.09808	0.02445	-0.02931	-0.059617	0.03375	-0.00590	-0.04294
3	15%	0.23124	-0.10706	0.05119	-0.03634	0.013910	0.09889	-0.07249	-0.24434
4	7%	0.15016	-0.16353	0.14590	0.70438	0.095090	0.20164	0.07417	0.01267
5	7%	0.26461	-0.12649	0.09048	-0.11265	-0.065167	0.00192	0.13361	-0.17256

Cluster	_SVD_9	_SVD_10	_SVD_11	_SVD_12	_SVD_13	_SVD_14	_SVD_15	_SVD_16	_SVD_17
1	-0.01451	-0.064338	0.03511	-0.00685	0.00120	0.039016	-0.025323	-0.012468	0.006683
2	-0.03920	0.030257	-0.04226	0.06938	-0.02547	-0.006180	-0.008381	0.008802	-0.021666
3	-0.15054	0.039647	-0.15791	-0.08368	0.14126	0.007295	0.074928	-0.037803	-0.008080
4	0.10782	0.026232	0.05786	-0.02929	0.01994	0.039150	-0.000047	-0.024032	0.045886
5	0.10158	-0.027458	-0.00909	-0.18089	0.08332	-0.024358	0.090719	-0.032576	-0.034352

Cluster	_SVD_18	_SVD_19	_SVD_20	_SVD_21	_SVD_22	_SVD_23	_SVD_24	_SVD_25	_SVD_26
1	0.009911	0.018363	-0.007464	0.000202	-0.010107	-0.001045	0.020362	0.004858	0.007114
2	-0.026244	-0.017876	-0.007149	0.027864	-0.002674	-0.014299	-0.032526	-0.008243	0.005134
3	0.012893	0.037562	0.050248	-0.005603	-0.013864	-0.002616	0.006280	0.032360	0.010095
4	0.007133	-0.034474	0.037352	-0.017737	0.010407	0.034208	-0.005033	0.010930	0.024545
5	-0.042197	0.011237	-0.051328	-0.043578	-0.080999	-0.019287	-0.013151	-0.014085	-0.017104

Cluster	_SVD_27	_SVD_28	_SVD_29	_SVD_30	_SVD_31	_SVD_32	_SVD_33	_SVD_34	_SVD_35
1	-0.008121	-0.015122	-0.007869	0.006945	-0.000802	-0.011948	-0.006151	0.004569	-0.022276
2	0.010833	-0.001224	0.011521	-0.001856	0.015333	0.001395	0.013645	-0.001673	-0.008444
3	-0.025280	0.042387	0.033086	0.010591	-0.006992	0.005913	0.009713	-0.043524	0.012099
4	-0.014677	-0.007524	0.012050	-0.006285	-0.031481	-0.006948	-0.004808	-0.010477	0.014757
5	0.009020	0.027118	0.038261	0.022961	0.010315	0.042643	-0.035409	0.012490	0.044207

Cluster	_SVD_36	_SVD_37	_SVD_38	_SVD_39	_SVD_40	_SVD_41	_SVD_42	_SVD_43	_SVD_44
1	0.016906	0.001971	0.009720	0.009558	0.006146	0.007855	-0.006398	0.002507	-0.006721
2	0.014133	0.010582	-0.012703	-0.000572	-0.011126	0.005253	0.008617	0.014142	-0.000722
3	0.036044	-0.007880	-0.032505	0.032204	-0.006572	0.000088	-0.043427	-0.015795	0.043941
4	-0.001950	-0.004910	-0.008525	-0.005799	-0.022377	-0.006062	-0.005111	-0.000615	-0.008600
5	-0.041415	0.015497	-0.018346	0.001734	0.026740	-0.013484	-0.021438	0.034649	0.013612

Cluster	_SVD_45	_SVD_46	_SVD_47	_SVD_48	_SVD_49	_SVD_50	_SVD_51	_SVD_52	_SVD_53
1	0.011748	-0.011308	0.006362	.00206003	-0.005140	.006565156	-0.004046	-0.008839	-0.007075
2	-0.009075	-0.018187	0.006534	.00227356	0.029812	.005920999	-0.017853	-0.014938	-0.015998
3	-0.002248	0.004548	0.016183	.00125867	-0.024166	.003912629	-0.041409	0.020351	-0.036183
4	0.020475	-0.008616	-0.002409	.00591111	0.002178	.008481086	0.014139	0.004928	-0.007321
5	0.031652	0.026516	-0.013068	.00026706	-0.010251	.007198722	0.007389	0.025615	-0.041482

Cluster	_SVD_54	_SVD_55	_SVD_56	_SVD_57	_SVD_58	_SVD_59	_SVD_60	_SVD_61	_SVD_62
1	-0.003995	-0.002159	0.009207	0.005231	-0.004804	.007856196	-0.002111	0.002017	-0.001542
2	0.012128	-0.016545	-0.016827	-0.006830	0.004662	.006545740	-0.027027	-0.014915	-0.008450
3	-0.014044	0.001561	0.002475	-0.022942	-0.002464	.007561171	-0.003771	-0.007483	0.006016
4	-0.000651	0.016312	0.007152	0.001499	-0.017239	.002346205	0.030272	0.007588	-0.001307
5	0.014084	-0.023110	0.021477	-0.002322	-0.006722	.004652785	0.006040	-0.020165	-0.011081

Cluster	_SVD_63	_SVD_64	_SVD_65	_SVD_66	_SVD_67	_SVD_68	_SVD_69	_SVD_70	_SVD_71
1	-0.011271	-0.002810	-0.003490	-0.005763	0.000454	.00481710	.00034602	-0.000356	0.001821
2	0.006102	0.008888	-0.003123	0.001675	-0.027728	.00187997	.00215860	-0.020886	0.006742
3	-0.021318	0.013377	0.022137	-0.025189	-0.014880	.00561440	.00132186	-0.019524	0.020367
4	-0.005084	0.008360	-0.023410	-0.007312	-0.003039	.00547972	.00385856	0.011434	-0.016902
5	-0.016684	-0.016267	-0.027122	-0.016775	-0.027949	.00415764	.00097843	0.004341	-0.005515

Cluster	_SVD_72	_SVD_73	_SVD_74	_SVD_75	_SVD_76	_SVD_77	_SVD_78	_SVD_79	_SVD_80
1	-0.012149	0.012534	0.003504	0.005793	-0.006698	0.002614	0.001364	-0.004862	0.007093
2	0.010177	0.021123	0.022107	-0.004343	-0.001668	0.008174	-0.014609	0.012519	-0.011828
3	-0.025713	0.034492	-0.000040	-0.009523	-0.019459	0.001780	0.014098	0.008011	-0.009752
4	-0.004123	-0.014506	-0.005869	0.002074	0.003446	-0.015116	0.004933	0.013661	-0.003142
5	-0.006910	-0.018475	0.003225	0.016491	-0.004907	0.018970	0.009586	0.006334	-0.000016

Cluster	_SVD_81	_SVD_82	_SVD_83	_SVD_84	_SVD_85	_SVD_86	_SVD_87	_SVD_88	_SVD_89
1	0.008000	-0.001248	-0.004539	0.000713	-0.008254	0.004335	-0.010372	).007502179	-0.001371
2	0.030755	0.010323	0.002959	-0.002599	0.006623	0.003703	0.001456	).002758810	-0.008381
3	0.021209	-0.001289	-0.001234	-0.029441	0.016281	0.004972	-0.011425	-.007858666	-0.002600
4	-0.011457	0.006109	-0.001401	-0.001606	-0.009569	-0.021730	-0.001627	-.000094409	-0.014802
5	-0.022116	0.017449	-0.020482	0.004515	0.009726	-0.017019	-0.010315	-.007051517	-0.005805

Cluster	_SVD_90	_SVD_91	_SVD_92	_SVD_93	_SVD_94	_SVD_95	_SVD_96	_SVD_97	_SVD_98
1	-.001637970	-0.008123	0.002028	-0.004197	-0.005084	-0.006881	-0.011381	-0.001911	0.004241
2	-.009890886	-0.008206	0.017911	0.010565	0.008142	0.006534	0.017607	-0.004866	0.002424
3	-.000644430	0.024254	0.003822	-0.023358	-0.001277	-0.026434	0.003631	-0.016148	0.025127
4	).000932414	0.000496	-0.003286	0.002418	-0.005416	-0.010896	-0.015728	0.009054	0.000093
5	).001562359	0.002378	-0.014937	-0.013732	-0.023945	0.005598	-0.007979	0.008929	0.009445

Cluster	_SVD_99	_SVD_100
1	-0.002616	-0.010367
2	-0.004818	-0.001567
3	-0.006928	0.006152
4	-0.006535	0.002976
5	-0.011869	0.015492

APÊNDICE D:

Transcrições de relatos representativos (Clusters Finais)

Transcrição 1:

“ SRA xxx ORIENTADA A RETORNAR LIG DE OUTRO PARELHO PARA REPR xxx E FEITA COM SUCESSO PARA APARELHO QUE NAO FAZ E RECEB xxx E LIG xxx 6 SRTA xxx INFORMA QUE NAO FAZ E NEM RECEBE CHAMADA xxx 7 MENSAGEM INFORMA QUE NAO ESTA DENTRO DO ENDERECO DE xxx 8 INSTALACAO xxx 9 FEITO CHECK LIST E DUAS REPROGRAMACOES SEM SUCESSO TRANFERITO PAR xxx 10 A TTT REPAROS xxx TTT ON LINE SRA xxx RECLAMA DE NAO RECEBER LIG ABERTO OR 30644140 13 7538311 CIENTE DE PROBLEMA NA REGIAO CIENTE DE CTT DA EMPRESA 30644140 14 CTT SRA xxx 30644140 15 TTT ON LINE 30644140”

Transcrição 2:

“OS 04919866 PONTO REF PROX AO METRO DE TOMAS COELHO, NA ASSOC xxx IACAO xxx 3 xxx 5 DT MASTERCARD xxx 6 ENDER ENTREGA xxx 39033407 8 RESP 1 SANDRA OU TIANE xxx 9 RESP xxx do Ciclo 13011824 foi processada para de xxx 12 Saldos e uso movidos do ciclo 17 para 02 para contas xxx Sra Deise pede inf sobre a entrega, a mesma esta na esta rlc,foi xxx aberta uma cmc cc, nada consta direct, nao tenho senha bslog xxx foi aberta uma cnv cc cliente inf prazo de tratamento xxx Ciclo 13022664 foi processada Saldos e uso movidos do ciclo 02 para 17 para contas TTT ABERTO PARA SRA DEISE NIT CC CIENTE DO PRAZO REPAROS SRA DEISE INF A RETORNAR DE OUTRO APARELHO PARA FAZER DO APARELHO SRA DEISE RECLAMA QUE FAZ E RECEBE LIG,FEITO CHECKLIST E , RESOLVEU PROBLEMA,TRANSF PARA TTT 28 foi infromado o numero do tel para sra deise,tel pago SR CARLSO INFORAMDO SOBER DATA DE VENMCIMENTO ADA FATURA SR DEISE FEITO MIGRACAO DE PLANO POS

55 P POS L 21 SOLICITACAO ATENDIDA FOI ENVIADO A FAT PARA O E MAIL DA SENHORA DEISE LANE xxx@gmail”

Transcrição 3:

“FEITO ENVIO DE 2VIA DE FATURA QUE VENCE HOJE A PEDIDO SR ADENIR P xxx 2 ARA E MAIL xxx.com br CLT INF SOBRE D A , SIGA ME E CONF A TRES 4 SR ADNIR ENCAMINHADO 2 VIA VALOR 55,05 VENCIMENTO 20 07 08 PELO xxx 5 E MAIL xxx@ oi com br ”

Transcrição 4:

“SRA LUCIENE CONF DADOS SOLICITA ENVIO DE 2 VIA DE FATURA xxx 3 FRA CLIENTE CONF DADOS E USO DA LINHA, SEM IND DE FRAUDE xxx 4 RETIRADO BLOQUEIO PREVENTIVO, PRAZO DE 30 24h xxx 5 PARA DA LINHA xxx 6 CLIENTE TEM HIST DE PGTO A MESMA DESEJA INF DE SUA FATURA TRANSF 7 P ATT 8 Srª Luciene solicita sobre 2 via de sua fatura pois xxx8 9 tyem email e o codigo de barras esta sendo aceito, client xxx 10 e tem fax aberto 2 via pelo correio REFT BIL xxx 11 39893058 12 SRA LUCILENE SOLICITOU DETALAHMENTO DE FATURA E xxx 13 SOBRE SALDO PARCIAL xxx 14 xxx 15 PREVENCAO CONSUMO xxx 16 Score I Limite 0 xxx 17 39893058 18 sr deivid solicitou fatura pelo seu email xxx 19 xxx yahoo com br xxx 20 enviado com sucesso xxx 21 SRA MARIA PEDE CADASTRO EM DEBITO AUTOM INF FAZER 1º NO BANCO, xxx 22 C 39893058 23 SR A xxx 24 INFORMADO A RETORNA COM DADOS EM xxx 25 SRA MARIA SOLICITOU A DE PLANO xxx 26 ABERTO UMA MIGP Nº Nº xxx 28 do Ciclo 13167660 foi processada para de xxx 29 Saldos e uso movidos do ciclo 01 para 26 para contas xxx Back Office Dispute xxx 32 BIL ATENDIDA xxx 33 Enviado a fatura ao conforme solicitado xxx 34 Back Office Dispute 39893058 35 sra maria ligou qurendo saber sobre fatura xxx”

Transcrição 5:

“A B 1 do Ciclo 13267828 foi processada para de 30641382 2 Saldos e uso movidos do ciclo 02 para 11 para contas 30641382 3 129872993 30641382 4 do Ciclo 13263023 foi processada para de 30641382 5 Saldos e uso movidos do ciclo 11 para 02 para contas 30641382 6 129872993 30641382 7 do Ciclo 13490985 foi processada para de 30641382 8 Saldos e uso movidos do ciclo 11 para 02 para contas 30641382 9 129872993 30641382 10 do Ciclo 13495732 foi processada para de 30641382 11 Saldos e uso movidos do ciclo 02 para 11 para contas 30641382 12 129872993 30641382 13 do Ciclo 13379730 foi processada para de 30641382 14 Saldos e uso movidos do ciclo 11 para 02 para contas 30641382 15 129872993 30641382 16 do Ciclo 13384982 foi processada para de 30641382 17 Saldos e uso movidos do ciclo 02 para 11 para contas..." *(demais trechos da transcrição suprimidos – pois os arquivos característicos deste Cluster são muito extensos)*

Transcrição 6:

"SR JOSE CARLOS SOLICITA DETALHAMENTO DE FATURA xxx 3 SR JOSE CARLOS ESTA COM FATURA PARA BOF ME V xxx 4 atendida, efetuado a troca de sem troca de xxx 5 xxx 07 6 Back Office Atendimento 34713907 7 SR JOSE CARLOS SOLICITA DE ABERTA B 34713907 8 OF ME 34713907 9 34713907 10 contato com sucesso 34713907 11 aprendiz 34713907"

APÊNDICE E:

Geração de uma lista de palavras semelhantes / sinônimos, baseada na Fonética dos Termos

```
%textsyn( termds=emws1.text2_terms,
          docds=emws1.text2_documents,
          outds=emws1.text2_out,
          textvar=text,
          mnparcdoc=8,
          mxchddoc=10,
          synds=50V2.vaerextsyns,
          dict=50V2.vaers.engedict,
          maxsped=15 );
```

APÊNDICE F:

Programação em XML / KML para Georeferenciamento de Pontos pelo Código de Endereçamento Postal (CEP) *

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

  <html xmlns="http://www.w3.org/1999/xhtml">

<head>

<script src="http://maps.google.com/maps?file=api&amp;v=2&amp;key= xxxxxxx "
type="text/javascript"></script>

  <!--chave do des01-->

  <!--xxxxxx-->

<script type="text/javascript">

var geocoder = new GClientGeocoder();

var map;

var ptr = 0;

var xml;

var markers;

var currcep;

void function Carrega(){

 GDownloadUrl("cep.xml", xmlprocessado);

}

function xmlprocessado(data, responseCode){

  xml = GXml.parse(data);

  markers = xml.documentElement.getElementsByTagName("ponto");

  currcep = markers[ptr].getAttribute("cep");

  geocoder.getLocations(currcep, GetNext);

}
```

```
function GetNext(response){
  place = response.Placemark[0];
  Lat = place.Point.coordinates[1];
  Long = place.Point.coordinates[0];
  var l="&lt;Placemark&gt;&lt;styleUrl&gt;#idstylecep&lt;/styleUrl&gt;&lt;name&gt;" + currcep + "&lt;/name&gt;&lt;Point&gt;&lt;coordinates&gt;" + Long + "," + Lat + "&lt;/coordinates&gt;&lt;/Point&gt;&lt;/Placemark&gt;";
  document.getElementById("map").innerHTML+=l+"<br>";
  ptr ++ ;
  if(ptr == markers.length){
    document.getElementById("map").innerHTML += "<hr> Fim ! <hr>";
    return;
  }
  currcep = markers[ptr].getAttribute("cep");
  geocoder.getLocations(currcep, GetNext);
}
</script>
</head>
<body onload="Carrega()">
<div id="map" style="width: 1500px; height: 1400px">
</div>
</body>
</html>
```

* As chaves (Keys) fornecidas pela Google para validação do Servidor em que os programas KML foram executados, estão por questão se Segurança da Informação, substituídas por “xxxxx”.

APÊNDICE G:

Estatísticas do Clusters Finais Obtidos

Distância entre Centróides_Clusters Finais

Cluster	1	2	3	4	5
1		0,957	1,088	0,955	1,105
2	0,957		0,536	0,241	0,627
3	1,088	0,536		0,528	0,754
4	0,955	0,241	0,528		0,620
5	1,105	0,627	0,754	0,620	

Final Cluster Centers

Variable	1	2	3	4	5
@_SVD_1	0,70663	0,05498	0,07376	0,05137	0,12548
@_SVD_2	0,51689	-0,0193	0,00388	-0,0197	0,05515
@_SVD_3	-0,2876	0,03988	0,00574	0,04311	0,00522
@_SVD_4	-0,0131	0,00553	0,00692	0,01851	-0,0112
@_SVD_5	0,05408	0,00097	-0,0267	-0,0128	-0,0176
@_SVD_6	-0,0396	-0,0076	0,00022	-0,0134	0,02284
@_SVD_7	-0,0391	-0,0011	0,02685	0,00294	0,02224
@_SVD_8	-0,0209	-0,012	-0,0254	-0,0407	-0,0013
@_SVD_9	0,01264	0,00649	0,01127	0,00169	0,03969
@_SVD_10	0,00181	0,01796	0,00169	-0,0107	-0,038
@_SVD_11	0,01643	-0,0336	-0,0313	-0,0198	-0,0334
@_SVD_12	0,05109	-0,0134	-0,0237	0,0263	-0,008
@_SVD_13	-0,0076	0,02414	0,00138	-0,0017	-0,0054
@_SVD_14	-0,0213	-0,0088	-0,0156	-0,0067	0,02555
@_SVD_15	0,01832	-0,002	0,00632	-0,0029	-0,0098
@_SVD_16	-0,0129	0,00117	0,00336	-0,0026	0,04336
@_SVD_17	0,02297	0,02156	0,04199	0,01365	-0,025
@_SVD_18	-0,0334	0,00152	-0,0217	-0,0051	0,08119
@_SVD_19	-0,0984	0,01047	0,01775	-0,0095	0,13618
@_SVD_20	-0,0539	-0,0036	0,03792	-0,0115	0,08425
@_SVD_21	0,01016	-0,0036	-0,0496	-0,0174	0,01789
@_SVD_22	0,06078	-0,0194	-0,0401	0,01138	-0,068
@_SVD_23	0,06561	-0,0173	-0,0414	-0,0015	-0,0777

@_SVD_24	0,00674	0,00567	0,02376	-0,0132	-0,0117
@_SVD_25	0,04256	-0,0095	-0,0011	-0,0055	-0,0415
@_SVD_26	-0,0083	0,01577	0,02573	-0,004	-0,2114
@_SVD_27	0,06018	-0,0045	0,03024	0,00907	-0,2019
@_SVD_28	0,07067	0,01682	-0,0126	0,00324	-0,0716
@_SVD_29	0,0275	0,01866	-0,0186	0,00418	-0,04
@_SVD_30	-0,0028	-0,0082	0,005	-0,0123	0,0316
@_SVD_31	-0,006	0,00397	0,03958	-0,0081	-0,0621
@_SVD_32	-0,0124	0,00175	0,00229	0,01658	0,04181
@_SVD_33	0,01854	-0,0122	0,00223	0,00037	-0,0548
@_SVD_34	0,02818	0,00343	0,00615	-0,0138	-0,1932
@_SVD_35	-0,0105	-0,0158	0,02693	-0,0037	0,11567
@_SVD_36	-0,0362	-0,0027	-0,0403	-0,0055	0,04045
@_SVD_37	-0,0247	0,00803	0,05036	-0,0113	-0,0638
@_SVD_38	0,0365	0,00882	-0,0718	-0,0139	0,0244
@_SVD_39	0,01528	0,01956	0,0074	0,01826	-0,11
@_SVD_40	0,03646	-0,0129	-0,0057	0,00232	-0,0883
@_SVD_41	-0,0103	0,01008	0,0227	0,00645	0,00408
@_SVD_42	-0,0294	-0,0064	0,04274	-0,0079	0,11442
@_SVD_43	0,00716	0,0014	0,04892	-0,0171	0,05708
@_SVD_44	-0,0188	0,01418	0,04704	-0,0093	0,01446
@_SVD_45	0,02072	-0,0042	-0,0753	-0,0114	0,00495
@_SVD_46	0,00761	-0,0096	0,01868	-0,0059	-0,0996
@_SVD_47	-0,0132	0,02259	-0,01	0,01343	-0,0211
@_SVD_48	-0,0102	0,01049	0,07731	-0,0107	0,0386
@_SVD_49	-0,0282	0,00457	0,05933	-0,0211	-0,0093
@_SVD_50	0,01895	-0,0052	-0,0201	0,00264	-0,081
@_SVD_51	0,01942	0,00952	-0,101	0,01382	-0,0809
@_SVD_52	0,01186	-0,0217	-0,0149	0,01415	0,00804
@_SVD_53	0,02499	0,00667	-0,0236	-0,0171	-0,0669
@_SVD_54	-0,0018	0,00397	0,03782	-0,0246	-0,0753
@_SVD_55	0,03678	-0,0027	-0,0381	0,00362	-0,0741
@_SVD_56	-0,0325	-0,0067	0,032	0,01853	0,06749
@_SVD_57	0,01701	-0,0126	-0,0833	0,00899	0,00289
@_SVD_58	-0,0079	-0,0258	-0,041	0,01109	-0,0097
@_SVD_59	0,01123	0,0091	-0,1097	-0,0189	0,0791
@_SVD_60	-0,0172	-0,0013	0,134	-0,0015	-0,0653
@_SVD_61	-0,0201	-0,0282	0,05652	0,01929	-0,0508
@_SVD_62	-0,0225	0,00316	-0,0623	0,00298	-0,0715
@_SVD_63	0,03041	0,04679	-0,1201	-0,0193	-0,0728
@_SVD_64	0,03534	0,01028	-0,0659	-0,0172	-0,022
@_SVD_65	-0,0119	0,01699	0,02599	-0,0157	-0,0114
@_SVD_66	0,007	-0,0025	0,04338	-0,012	0,02909
@_SVD_67	-0,0011	0,02041	0,08008	-0,0157	-0,0055

@_SVD_68	0,0272	0,01601	0,00933	-0,0136	-0,0256
@_SVD_69	0,01891	-0,0161	-0,0501	0,0165	0,0184
@_SVD_70	0,02641	0,00533	-0,0407	0,0046	-0,0095
@_SVD_71	0,01832	-0,0042	-0,0264	-0,0008	-0,0432
@_SVD_72	0,02547	0,00268	-0,0452	-0,0112	-0,0671
@_SVD_73	0,01456	0,00739	-0,0496	0,00935	0,01367
@_SVD_74	-0,0288	-0,0107	0,10795	0,00781	0,04902
@_SVD_75	-0,0065	-0,0226	0,0128	0,01191	-0,0311
@_SVD_76	-0,0343	-0,0119	0,00404	0,02199	-0,0125
@_SVD_77	-0,0154	0,0236	0,06146	-0,0129	0,00267
@_SVD_78	-0,0176	-0,0032	0,01147	0,00046	-0,0134
@_SVD_79	-0,0157	-0,0334	0,05943	0,0307	0,04246
@_SVD_80	0,01031	-0,002	0,02049	-0,0005	0,02304
@_SVD_81	0,00184	-0,0085	0,01916	-0,0031	0,00203
@_SVD_82	0,00682	-0,0053	-0,0318	0,01838	0,01215
@_SVD_83	0,03353	-0,0108	-0,0379	0,01859	0,02043
@_SVD_84	0,0185	0,01696	-0,0355	-0,0137	-0,007
@_SVD_85	0,01743	-0,0475	0,04189	0,03715	-0,0409
@_SVD_86	0,0038	-0,0091	-0,007	-0,0002	0,00257
@_SVD_87	0,01289	-0,0026	-0,0649	0,01884	-0,0343
@_SVD_88	-0,0259	0,01167	0,14818	-0,0047	0,00328
@_SVD_89	0,02368	-0,0138	-0,0494	-0,0066	0,03453
@_SVD_90	-0,0046	0,00256	-0,0866	0,01034	-0,0355
@_SVD_91	-0,0053	-0,028	0,04742	0,01483	0,01167
@_SVD_92	0,0156	-0,026	-0,0484	0,03681	-0,0204
@_SVD_93	-0,0008	-0,0132	0,07058	0,01612	0,00711
@_SVD_94	-0,028	-0,0188	0,04195	0,01485	0,03801
@_SVD_95	0,00372	0,00731	-0,0576	0,0043	-0,0108
@_SVD_96	-0,0133	0,00541	-0,108	-0,0044	0,02783
@_SVD_97	-0,0132	0,01015	0,06525	-0,0036	0,00153
@_SVD_98	0,01277	-0,0167	-0,0413	0,01685	-0,0323
@_SVD_99	0,0276	-0,0003	0,11175	0,00494	-0,0404
@_SVD_100	-0,0005	0,00722	0,05402	0,00575	-0,0002