UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ

CAMPUS LUIZ MENEGHEL

RAFAEL APARECIDO LOPES

MINERAÇÃO DE TEXTOS

APLICADA A UMA BASE DE RECLAMAÇÕES SOBRE PRODUTOS E SERVIÇOS

Bandeirantes

2009

RAFAEL APARECIDO LOPES

MINERAÇÃO DE TEXTOS
APLICADA A UMA BASE DE RECLAMAÇÕES SOBRE PRODUTOS E SERVIÇOS

Trabalho de Conclusão de Curso submetido à Universidade Estadual do Norte do Paraná – Campus Luiz Meneghel, como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Msc. Glauco Carlos Silva

Bandeirantes

2009

RAFAEL APARECIDO LOPES

MINERAÇÃO DE TEXTOS
APLICADA A UMA BASE DE RECLAMAÇÕES SOBRE PRODUTOS E SERVIÇOS

Trabalho de Conclusão de Curso submetido à Universidade Estadual do Norte do Paraná – Campus Luiz Meneghel, como requisito parcial para a obtenção do grau de Bacharel em Sistemas de Informação.

COMISSÃO EXAMINADORA

_____

Prof. Msc. Glauco Carlos Silva
UENP – Campus Luiz Meneghel

_____

Prof. Msc. Ailton Sérgio Bonifácio
UENP – Campus Luiz Meneghel

_____

Prof. Dra. Marilia Abrahão Amaral
UENP – Campus Luiz Meneghel

Bandeirantes, __ de _____ de 2009

A Deus, aos meus pais, aos professores e aos meus amigos...

companheiros de todas as horas...

AGRADECIMENTOS

Primeiramente a Deus, que é lâmpada para os meus pés e luz para o meu caminho.

A minha família, pela confiança, motivação, incentivo e apoio durante todos os momentos da minha vida, principalmente durante a realização deste trabalho.

Ao Prof. Msc. Glauco Carlos Silva, braço amigo no desenvolvimento deste trabalho, que sempre me orientou com muita inteligência e clareza, norteando os passos a seguir em cada etapa deste projeto.

Ao amigo Luiz Gustavo Detone Romão, pela colaboração com idéias, ferramentas e todo apoio provido durante a realização do trabalho.

Aos amigos e colegas da X turma de Sistemas de Informação da UENP/CLM, Kairo, Gustavo, Danilo, André, Carlos, Marcel e Paula, pelos momentos de apoio, companheirismo, incentivo e descontração. Das horas de desespero às horas de alegria, amigos para sempre.

Aos professores, pelos momentos de pressão, descontração, aconselhamento e discussões, que fizeram com suas boas intenções, mesmo que nos recusássemos a enxergar. Levo ótimas lembranças de todos, dos mais extrovertidos aos mais sérios, dos mais espontâneos aos mais reservados, levo as lembranças para a vida.

Ao amigo Leandro Krug Wives, pela disponibilização da ferramenta e dicas importantes sobre o trabalho.

Ao amigo Normann Kalmus, pela disponibilização da base de documentos, peça fundamental para a realização da pesquisa.

A todos que, com boa intenção, colaboraram para a realização e finalização deste trabalho.

“Avalia-se a inteligência de um indivíduo pela quantidade de incertezas que ele é capaz de suportar.”

Immanuel Kant

Resumo

A informação vem exercendo um papel essencial no desenvolvimento das grandes empresas. Com isso o volume dados gerados pelas organizações tem aumentado consideravelmente. Estes dados, na maioria das vezes não possuem estruturas, com isso ocultam conhecimento que poderia ser de grande utilidade para as organizações. Neste trabalho foi abordado tarefas de Mineração de Textos. Foram descritas todas as etapas do processo, desde a coleta de dados até a avaliação dos resultados. Para tanto, foi utilizado uma base de reclamações de produtos e serviços, na qual foram aplicadas as etapas de mineração. Por final, foi realizado um exemplo de categorização automática, onde foram aplicados dois algoritmos específicos para criação de agrupamentos automáticos e em seguida foram comparados os resultados.

Palavras-chave: Mineração de Textos, Descoberta de Conhecimento em Textos, Inteligência Artificial.

Abstract

Information is playing a key role in the development of large enterprises. Thus the volume data generated by organizations has increased considerably. These data, in most cases do not have structures, thereby concealing knowledge that could be very useful for organizations. This work was addressed tasks of Text Mining. Were described every step of the process from data collection to the evaluation of the results. To this end, we used a basis for claims of products and services, which were applied in steps of mining. By the end, was an accomplished example of automatic categorization, applying two specific algorithms for automatic clustering and then compared the results.

Keywords: Text Mining, Knowledge Discovery in Texts, Artificial Intelligence.

LISTA DE ABREVIATURAS

AM – Aprendizado de Máquina

DCD – Descoberta de Conhecimento em Dados

DCT – Descoberta de Conhecimento em Textos

IE – Information Extraction

KDD - Knowledge Discovery in Data

KDT – Knowledge Discovery in Texts

MD – Mineração de Dados

MT – Mineração de Textos

PLN – Processamento de Linguagem Natural

SAC – Serviço de Atendimento ao Consumidor

TF – Term Frequency

TXT – Documento de Texto Puro

WEKA – Waikato Environment for Knowledge Analysis

XML – Extensible Markup Language

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 Introdução

Atualmente, empresas e pessoas produzem uma ampla quantidade de documentos eletrônicos. Grandes coleções de documentos textuais são elaboradas a cada dia, várias novas páginas são lançadas na web, formando um conglomerado de documentos (REZENDE, 2005).

Diante desta constatação, surgiu a possibilidade de se aplicar técnicas capazes de valorizar estes dados. A Mineração de Textos se enquadra neste contexto, visto que, é um conjunto de técnicas e processos que descobrem conhecimento inovador nos textos (REZENDE, 2005). Trata se de um conjunto de técnicas que abordam dados não estruturados ou semi estruturados. A aplicação de Mineração de Textos pode ser aplicada para uma simples classificação de documentos ou para descobrir conhecimentos e auxiliar gerentes na tomadas de decisões.

Hoje em dia, estão disponíveis recursos computacionais que possibilitam acesso à informação de maneira rápida e eficiente, desde que a mesma esteja organizada em banco de dados apropriados à manipulação por computadores. Grande parte da informação eletrônica encontra-se disponível em bases de dados freqüentemente chamadas de não-estruturada, ou seja, bases de documentos textuais, cujo formato está adequado ao homem que, através da leitura, é capaz de decodificar a informação contida no texto e aprendê-la (SCHIESSL, 2007). Por outro lado nestes documentos há informações que o homem não consegue absorver, dessa maneira, a máquina desempenha um papel fundamental na gestão da informação.

A Descoberta de Conhecimentos em Textos (DCT), através da ajuda de máquinas, tem a finalidade de propor soluções para tratar a informação eletrônica em formato textual, a fim de diminuir o impacto da carga excessiva de informação (RINO e PARDO, 2003).

1.1 Objetivos

1.1.1 Objetivo geral

Estudar e utilizar as técnicas de Mineração de Textos através da utilização de ferramenta, que receba como entrada arquivos em formato TXT, aplicando-a na base de reclamações sobre produtos e serviços. Estes dados pertencem a uma iniciativa de uma empresa privada que disponibiliza este serviço. Espera-se ainda ao longo das conclusões das etapas de Mineração, descobrir novos conhecimentos na base, informações não triviais, e adquirir experiência sobre cada etapa do processo de Mineração de Textos.

1.1.2 Objetivos específicos

Os objetivos específicos do projeto se resumem em:

- Estudar as técnicas de Mineração de Textos;
- Aplicar as técnicas de Mineração de Textos:
  - Coleta de Dados;
  - Pré-Processamento;
  - Extração de Padrões;
  - Avaliação e Interpretação;
- Extrair Conhecimento da Base de Reclamações com aplicação de DCT e identificar e/ou propor categorias de agrupamento com base no conteúdo das reclamações.
- Criar agrupamento para os documentos, de modo que os documentos pertencentes a certo agrupamento possuam baixo grau de divergência entre si e alto grau de diferença em relação a documentos de outro grupo distinto.

1.2 Justificativa

Visto a enorme disponibilidade de dados eletrônicos, surge a necessidade de extrair conhecimentos destes dados. Conhecimento este, que é impossibilitado de ser visualizado sem tais recursos de mineração. Através de técnicas de mineração de textos, é possível realizar a extração de informações úteis capazes de serem compreendidas por humanos. É esperado que, através da mineração de textos a manipulação perfeita destes seja alcançada, possibilitando classificá-lo, categorizá-lo, enfim, realizar automaticamente várias tarefas que antes só eram possíveis de serem realizadas humanamente.

A escolha da base de reclamações é dada pelo fato do conteúdo ser amplo e abordar assuntos distintos, podendo chegar a resultados valiosos e interessantes, que podem comprovar a eficiência de cada passo do processo de mineração de textos. Ainda é esperada a possibilidade de encontrar informações intrínsecas e indicadores que permitam traçar o nível e as principais áreas reclamadas pela comunidade de consumidores.

Computadores não conseguem entender documentos textuais formatados para o entendimento de seres humanos. Os documentos, em seu estado natural, necessitam de pré-processamento antes de sua manipulação ou mineração computadorizada para a descoberta de padrões e relacionamentos entre os documentos da coleção. Embora a mente humana reconheça capítulos, parágrafos e sentenças, os computadores requerem dados que estejam organizados na forma de matrizes com linha, colunas e contagem de freqüências (SCHIESSL, 2007).

Dada a estruturas dos dados, a pesquisa em descoberta de conhecimento é uma necessidade da ciência e, conseqüentemente, as organizações possam se beneficiar da grande quantidade de informação potencialmente útil e desconhecida.

É justificada também pela carência de pesquisas direcionadas a esta área. Por ser uma área nova, a quantidade de pessoas capacitadas no assunto é pequena, com isso, toda a iniciativa focada nesta linha é muito bem aceita e apoiada pelos que já estão pesquisando.

2 Mineração de Textos

Assim como as técnicas de Mineração de Dados foram desenvolvidas para dados estruturados, as técnicas de Mineração de Textos focam o processamento de dados sem estruturação.

2.1 O que é Mineração de Textos

Hoje com o aumento de documentos eletrônicos, há um crescimento exorbitante na pesquisa de Mineração de Textos, bem como na criação de aplicativos que apresentem melhores alternativas para um bom aproveitamento destes documentos (MARTINS, 2003).

Segundo Rezende (2005), Mineração de Textos é um conjunto de técnicas e processos que descobrem conhecimento inovador nos textos.

Mineração de dados textuais, ou simplesmente Mineração de Textos, é o processo utilizado para descobrir padrões interessantes e úteis em um conjunto de dados textuais (MARTINS, 2003).

Para Tan (1999), sabendo-se que grande parte das informações das empresas estão contidas em documentos textuais, Mineração de Textos, é um grande desafio e uma tarefa complexa, já que trata dados sem estrutura. Define-se ainda, Mineração de Textos como sendo um campo multidisciplinar, envolvendo recuperação de informação, analise de texto, extração de informação, agregação, classificação, visualização, tecnologia de banco de dados, aprendizagem automática e mineração de dados (MOONEY & BUNESCU, 2005).

Conforme (Rajman e Besançon, 1997), bases textuais se apresentam de forma não estruturada. Porém, possuem uma estrutura implícita que necessita de técnicas específicas para ser reconhecida por sistemas automatizados. O Processamento de Linguagem Natural (PLN) trata exatamente da descoberta destas estruturas implícitas, como por exemplo, a estrutura sintática.

2.2 Tarefas de Mineração de Textos

Cada tipo de tarefa extrai um tipo diferente de informação dos textos.

Indexação: permite a procura eficiente em textos por documentos relevantes a *query* sem precisar examinar os documentos inteiros. Os tipos mais comuns de indexação são a indexação do texto completo e a indexação temática. Existe também a indexação tradicional, a indexação por *tags* e indexação semântica latente (REZENDE, 2005).

Categorização: induz um classificador que possa determinar se o documento pertence ou não a uma categoria pré-definida. A categorização é uma tarefa bastante utilizada no caso de existir um conjunto de documentos previamente classificados (MARTINS, 2003).

Agrupamento: ou *clustering* de documentos, consiste em agrupar os documentos em um conjunto finito de *clusters*. Técnicas de agrupamento de textos levam em conta as palavras que aparecem nos documentos para definir a função de similaridade e a fim determinar o agrupamento final (MARTINS, 2003).

Sumarização: Segundo Rino e Pardo (2003), a sumarização reduz o documento sem perder seu significado. Já Martins (2003) complementa que, o sumário de um documento é derivado de um texto fonte condensado pela seleção e/ou generalização em palavras ou frases chaves presentes em textos o qual mantém o conteúdo informativo do documento original.

2.3 Etapas de Mineração de Textos

Para a realização das tarefas, algumas etapas são comuns para a maioria delas (Rezende; 2005). As etapas são ilustradas conforme a Figura 1:

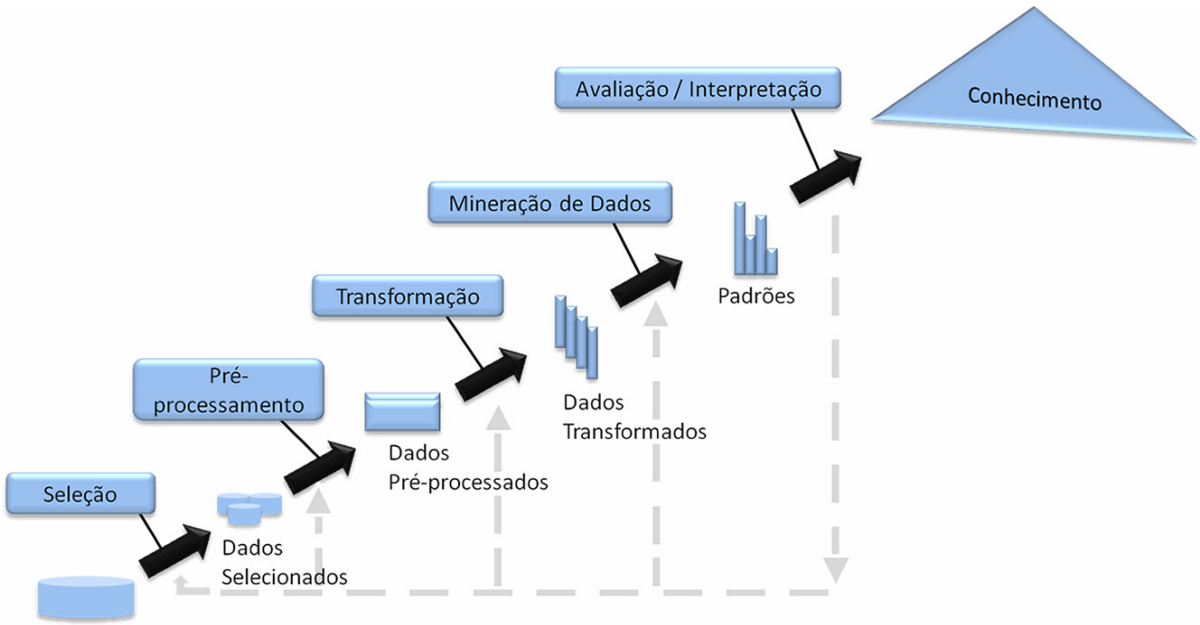

Figura 1 Etapas de Mineração de Textos (Fonte: Fayyad et al, 1997, apud, Schiessl, 2007).

- Seleção: consiste em recuperar documentos relevantes ao domínio de aplicação do conhecimento a ser extraído. Diversas fontes estão disponíveis para coletar documentos, tais como livros e, principalmente, documentos provenientes da Internet.

- Pré-processamento: é responsável pela limpeza da base. Realiza a eliminação de termos freqüentes, atividades de *stemming*, registros duplicados, palavras erradas, entre outras ações a fim de melhorar e reduzir a dimensionalidade do corpus.

- Transformação: esta etapa é responsável por transformar os documentos coletados, em um formato propício para serem submetidos aos algoritmos de extração de conhecimento. Nesta etapa obtém-se uma representação estruturada dos documentos, geralmente no formato de uma tabela atributo-valor.

- Mineração de Dados: tem por objetivo descobrir padrões úteis e desconhecidos presentes nos documentos. Caso os documentos estejam representados no formato de uma tabela atributo-valor, geralmente, são empregados métodos de Recuperação de Informação (RI) e de Aprendizagem de Máquina (AM) para a extração de padrões de forma semelhante ao processo tradicional de MD (Mineração de Dados).

- Avaliação e Interpretação: Essa etapa é necessária para verificar se o objetivo foi alcançado ou se necessitam ser refeitas todas ou algumas das

etapas do processo. A verificação pode ser feita pelo usuário final, especialista do domínio e/ou analista de dados, e então os resultados obtidos são analisados para verificar se condizem com o objetivo esperado.

2.4 Extração de Informação

Segundo Mooney e Bunescu (2005), infelizmente em muitas aplicações, as informações eletrônicas estão disponíveis sob a forma de documentos em linguagem natural não estruturada e em base de dados não estruturadas. Conseqüentemente, a Mineração de Texto, ou seja, a descoberta de conhecimentos úteis em textos não estruturado está se tornando cada vez mais um aspecto importante do KDD (*Knowledge Discovery in Data*).

Extração de Informação (IE) diz respeito a localizar dados em documentos de linguagem natural, assim, extrair informação estruturada a partir de dados não estruturados. Um tipo de IE, denominado entidade-reconhecimento, envolve a identificação referente a determinados tipos de objetos, tais como pessoas, empresas e locais. Além de reconhecer entidades, um problema importante é extrair tipos específicos de relacionamentos entre entidades (MOONEY e BUNESCU, 2005).

2.5 Análise Estatística

Há duas formas de abordagem dos dados. A análise semântica baseada em funcionalidade dos termos nos textos e a análise estatística, baseada em freqüência. Esses tipos de abordagem podem ser utilizados sozinhos ou em conjunto para análise dos dados.

Neste trabalho optou-se por utilizar a Análise Estatística na abordagem aos dados.

Na Análise Estatística, a importância dos dados é dada basicamente pelo número de vezes que eles aparecem nos textos. É interessante ressaltar que este tipo de estratégia pode ser conduzido independentemente do idioma (REZENDE, 2005).

Passos do aprendizado estatístico (REZENDE, 2005):

- Codificação dos Dados: Uma codificação inicial dos dados é escolhida ou baseada nas indicações de um especialista. Faz-se a seleção de

características informativas. Se informações relevantes forem descartadas nesse estágio, não poderão ser recuperadas mais tarde. Por outro lado, se conter muitas informações irrelevantes ou ruído, a procura por um bom modelo se torna difícil ou consome muito tempo, e as propriedades dos dados podem ser perdidas em meio ao ruído.

- Estimativa dos Dados: Um algoritmo ou um modelo de estimativa é aplicado para obter um modelo para os dados, normalmente pela maximização e algum critério. Este estágio pode ser visto como a procura por um modelo adequado a partir de um amplo conjunto de modelos possíveis.

- Modelos de Representação de Documentos: representam a idéia de saco de palavras (*bag of words*), que ignoram a ordem das palavras assim como qualquer informação de pontuação ou estrutural, mas retém o número de vezes que uma palavra aparece. Embora não seja elaborada o suficiente para uma interpretação completa da linguagem, a codificação *bag of words* provê uma quantidade considerável de informações sobre associações entre palavras e documentos.

2.6 Compartilhando Informações

Ainda sobre Mineração de Textos (MT), Martins (2003) afirma que:

> *"Embora, MT tenha um potencial comercial maior do que o próprio processo de MD, é um processo muito mais complexo que MD, pois trabalha com dados que são textos, ou documentos, que são inerentemente não estruturados e que, muitas vezes, possuem ambigüidade."*

Para Martins (2003) as complexidades da tarefa de MT estão relacionadas a vários fatores, tais como:

- A língua na qual o documento está escrito. Os algoritmos que manipulam textos são, geralmente, dependentes da língua.

- O estilo no qual o documento está escrito. Por exemplo, documentos mais formais são mais fáceis de serem processados do que documentos informais;

- A natureza do conteúdo do documento. Freqüentemente, documentos grandes que contém bastante informação irrelevante são difíceis de

serem processados, assim como documentos que contem informação não textual, como figuras;

- A especificação da tarefa a ser realizada, a qual depende do nível de detalhe e natureza da informação buscada, bem como a adequação das estruturas internas escolhidas para representar a informação.

3 O Serviço de Reclamações Online

As empresas nem sempre focaram a satisfação do cliente ao oferecer seus produtos e serviços. (Kotler,1972, *apud*, Schiessl, 2007) apontou a satisfação do cliente como elemento fundamental, na teoria do Marketing, para a sustentabilidade das organizações.

No Brasil, vive-se um curto período de proteção garantida pelo estado aos clientes das empresas. Nesse aspecto, aconteceu uma nova orientação com foco no consumidor e na sua satisfação garantindo a fidelização dos clientes, sustentando assim a lucratividade em longo prazo (KOTLER,1972, *apud*, SCHIESSL, 2007).

Recebendo um novo destaque no mercado o cliente ganhou espaços e meios para expressar sua satisfação ou frustração em relações aos produtos e serviços das empresas. As empresas, por sua vez, podem utilizar-se destas manifestações vindas dos consumidores pra realinhar seus produtos e serviços com foco no cliente e reforçar o vínculo cliente-empresa.

Com essa mentalidade atual, muitas empresas passaram a incentivar o consumidor a manifestar sua opinião em relação às organizações. Diante desta nova postura empresarial, iniciaram a criação de canais de comunicação que ajuda as empresas a corrigirem produtos, serviços e até mesmos a sua própria estratégia dentro do mercado.

Como formas de comunicação entre empresa e cliente surgiram os Serviços de Atendimento ao Consumidor (SAC). Muitas empresas adotam este meio de comunicação por considerarem valiosas as opiniões de seus clientes. Barlow e Moller (1996) colocaram que as reclamações são um presente ao planejamento estratégico das empresas. Ainda, elas se constituem de informações de baixo custo e sem intermediários no processo de comunicação cliente-empresa.

Como opção paralela aos SACs, existe implantado um portal online onde o consumidor pode exercer sua cidadania expressando sua reclamação quanto a atendimento, compra, venda, produtos e serviços. Este serviço possui a vantagem de permitir a reclamação contra qualquer empresa ou produto, e também consentir o direito de resposta por parte do destinatário da queixa.

Trata-se de serviço sem qualquer custo, que veio sanar a dificuldade de comunicação com as empresas que não possuem a implementação de um Serviço de Atendimento ao Consumidor (SAC) exclusivo, principalmente pequenas e médias empresas.

Enfim, em tempo de negócios com concorrência assídua, os consumidores desejam que seus nomes sejam conhecidos e que suas diferenças sejam encontradas e respeitadas pelas empresas, oferecendo assim, um atendimento personalizado. Com tanta informação direta do consumidor, tratando-as, as empresas podem extrair conhecimentos valiosos das mesmas.

4 Metodologia

O trabalho foi realizado através de pesquisas bibliográficas relacionadas à importância e utilidade de tratar documentos textuais e, a partir destes, extrair conhecimento. Também foi realizada pesquisa aplicada, com o objetivo de gerar conhecimentos para a aplicação prática. Serão mostradas como estão às pesquisas sobre o assunto atualmente e onde as grandes empresas do ramo pretendem chegar. Os estudos foram feitos com base em livros, artigos publicados, sites e revistas que destacam como vem sendo utilizadas estas técnicas.

4.1 Embasamentos

A mineração de texto pode possuir uma abordagem voltada tanto para a compreensão do conteúdo, quanto para a análise estrutural de documentos, isto é, análises estatísticas, e ambas objetivam a identificação de padrões implícitos em uma ampla quantidade de documentos.

Nesta perspectiva, é de grande importância explorar essas duas maneiras de analise e, para tanto, alguns conceitos fundamentais são indispensáveis para o aprofundamento do tema:

- *Token* é uma seqüência contígua de caracteres que não possuem um separador. Um separador é um caractere especial tal como um espaço em branco ou sinal de pontuação.
- Termo é um *token*, ou mais, com significado específico em uma determinada linguagem.
- Documento incide em um conjunto de *token.*
- Corpus é uma coleção de documentos.

Esses são os fundamentos básicos para iniciar um trabalho de mineração de textos para qualquer finalidade. Vale lembrar que neste trabalho, *Token* e Termo serão tratados como fundamentos iguais.

4.2 Recursos Utilizados

Os recursos utilizados para a execução do projeto foram:

Ferramentas:

- Microsoft Windows XP Professional service Pack 3 – Sistema Operacional;
- Pretext 2: ferramenta para pré processamento de textos;
- Eurekha: ferramenta de mineração de texto (algoritmo Star);
- WEKA: ferramenta de mineração de texto (algoritmo k-means);
- Intext: separação da base em arquivos textos individuais;
- TextPad: manipulação de arquivos;
- Microsoft Excel: criação de tabelas e gráficos.

Equipamento:

- PC AMD Turion 64 X2 com 2Gb de memória RAM e 160 Gb de disco.

Dados:

Os dados provêm de uma base de reclamações, onde usuários enviam mensagens reclamando sobre produtos e serviços.

As mensagens foram disponibilizadas em formato de planilha .xls, e cada linha representava um documento.

4.3 Ferramentas

Para o desenvolvendo deste trabalho foram utilizadas três ferramentas, sendo uma somente para pré-processamento e as outras duas para aplicação dos algoritmos de mineração.

4.3.1 Pretext

A ferramenta foi desenvolvida pelo Insititudo de Ciências Matemáticas e de Computação da USP e faz o pré-processamento dos textos. Escolheu-se essa ferramenta pelo fato de receber arquivos textos como entrada e por ser uma ferramenta consolidada para este tipo de tarefa. Por intermédio do Pretext, consegue-se fazer a *tokenização*, remoção de *stopwords*, a formação de *stemming* e se necessário, pode-se realizar o corte de palavras baseado na freqüência.

Como saída são gerados os arquivos discover.names e discover.data.

O arquivo discover.names: este arquivo contém a declaração de todos os atributos da tabela atributo-valor gerados pelo Pretext, como é ilustrado no Quadro 1.

Quadro 1: Discover.names: Atributos gerados pelo Pretext.

```
filename:string:ignore.
"polit":integer.
"regul":integer.
"proced":integer.
"tranquil":integer.
"valoriz":integer.
```

Quadro 2: Discover.data: Valores dos atributos.

```
"exemplo_Mail/exemplo1.txt",6,0,0,0,0
"exemplo_Mail/exemplo2.txt",1,3,0,0,0
"exemplo_Mail/exemplo3.txt",1,1,1,0,0
"exemplo_Mail/exemplo4.txt",1,1,1,1,0
"exemplo_Mail/exemplo5.txt",5,8,7,5,9
```

O arquivo *discover.data*: neste arquivo estão os valores dos atributos para todos os documentos da coleção, e representa a tabela atributo-valor como mostrado no Quadro 2.

. O primeiro atributo de todas as tabelas geradas pelo Pretext é sempre o *filename* que representa o documento que originou esses valores para seus atributos. Geralmente, este atributo não é utilizado no aprendizado de máquina, portanto é ignorado**.** A segunda linha do arquivo *discover.names* (Quadro 1) corresponde ao atributo cujo o valor está representado na primeira coluna do arquivo *discover.data* (Quadro 2), a terceira linha corresponde ao segundo atributo, e assim por diante.

4.3.2 WEKA

Para a tarefa de extração de conhecimento, o escolheu-se o software WEKA – *Waikato Environment for Knowledge*. Tal escolha ocorreu pela

razão de ser um programa livre e muito utilizado em vários trabalhos científicos, decorrendo em bons resultados.

O WEKA está implementado na linguagem Java, por isso possui portabilidade. Desta maneira, consegue ser executado nas mais variadas plataformas, aproveitando as vantagens de uma linguagem orientada a objetos como modularidade, polimorfismo, encapsulamento, reutilização de código dentre outros. Além disso, é um software de domínio publico estando disponível em http://www.cs.waikato.ac.nz/ml/weka/ (SANTOS, 2005).

De acordo com Santos (2005), o WEKA aplica manipulação em um arquivo com extensão .ARFF que possui texto puro e é formado por três partes:

- Relação – no inicio do arquivo, especificamente na primeira linha, deve conter a identificação da tarefa que esta sendo realizada, sendo antecedida pela expressão @relation;
- Atributos – abaixo da linha de identificação, deve-se relacionar os atributos, sendo que em cada linha coloca-se a expressão @attribute seguido do nome do atributo e seu respectivo tipo;
- Dados – no final, deve descrever os dados colocando a expressão @data e abaixo representar em cada linha uma instância dos dados separando seus atributos por vírgula.

O arquivo pode ter linhas de comentários que não recebem processamento. Para inserir comentários, deve-se iniciar cada linha com o sinal de porcentagem (%).

Quadro 3: Estrutura do Arquivo .ARFF

```
@relation ensaio

@attribute docs {T0001,T0002,T0003,T0004,T0005,T0006,T0007}
@attribute recarg {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute acion {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute administra {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute marketing {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute venc {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute transit {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute observan {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
```

```
@data
T0001,0,10,0,1,0,3,0
T0002,0,0,0,0,14,0,0
T0003,0,0,0,11,0,0,0
T0004,0,3,0,0,0,11,0
T0005,0,7,0,0,12,0,0
T0006,0,0,12,0,1,0,0
T0007,0,0,0,0,0,15,0
```

4.3.3 EUREKHA

É uma ferramenta que permite ao usuário uma interatividade durante o processo de obtenção de padrões e relacionamento. Foi implementada utilizando a linguagem C++, com algumas características de orientação a objetos.

Tem o objetivo de auxiliar no processo de exame e recuperação de informações de base de dados textuais. Permite ainda a extração de conhecimentos de base não estruturada.

A ferramenta dispõe de uma série de alternativas para a análise de um corpus. Dentre estas alternativas está a análise lexical, onde são listadas as palavras pertencentes a cada documento e seus respectivos valores de freqüência e relevância, e a análise de centróide, que gera um gráfico contendo as palavras mais relevantes de determinado agrupamento, facilitando a identificação do assunto por ele representado (Wives, 1999).

4.4 Algoritmos

4.4.1 Stars

Este algoritmo consiste em selecionar um elemento e identificar todos os elementos conectados a ele. Deste modo, tem-se uma figura muito parecida com uma estrela (daí o nome: *star* ou estrela), pois um item central conecta todos os outros componentes do grupo.

1) Selecionar 1 termo e colocar todos os similares na mesma classe;

2) Termos ainda não classificados são colocados como semente de classe

Porém, ele apresenta dois problemas. O primeiro refere-se ao fato dos objetos serem alocados no primeiro grupo onde o grau de similaridade seja

maior do que o mínimo estipulado. E, já que os objetos alocados em um grupo não são mais processados, não há garantia de que um objeto tenha sido colocado no grupo de maior afinidade (similaridade). Além disso, caso a ordem dos elementos na matriz de similaridades seja trocada, o resultado do agrupamento pode variar.

Outro problema encontra-se no fato que processo não indica todos os grupos que o objeto poderia fazer parte. Em documentos é possível que existam textos que tratem de mais de um assunto. Utilizando ao algoritmo *stars,* o documento seria atribuído ao primeiro assunto que atendesse a restrição do grau mínimo de similaridade (WIVES, 1999).

4.4.2 K-means

O algoritmo *k-means* é um dos mais simples algoritmos de aprendizado não supervisionado que resolve o problema de segmentação por agrupamento.

De modo geral, esse processo incide em usar os valores dos primeiros n casos em um arquivo de dados, como estimativas temporárias das médias dos k conjuntos, onde k é o número de conjuntos apontado pelo usuário. Assim, o centro do conjunto inicial, chamado de centróide, é criado para cada caso em torno dos dados mais próximos e depois comparados com os pontos mais distantes e os outros conjuntos formados. Após essa etapa, ocorre um processo de atualização contínua e interativa, encontrando os centros dos conjuntos finais ao término do processo (SANTANA, 2008).

De uma forma prática e ilustrada, será apresentado passo a passo como o k–means funciona:

1. Aleatoriamente são gerados 3 centróides (k =3);
2. Atribuir a cada um dos objetos o conjunto que tem o centróide mais próximo
3. As posições dos centróides são recalculadas (nessa etapa nota-se que alguns pontos são deslocados de acordo com o resultado);
4. Se a posição dos centróides não mudar, passa-se para a próxima etapa. Se não, voltará para a segunda etapa.
5. Resultado do processo, com todos os pontos agrupados.

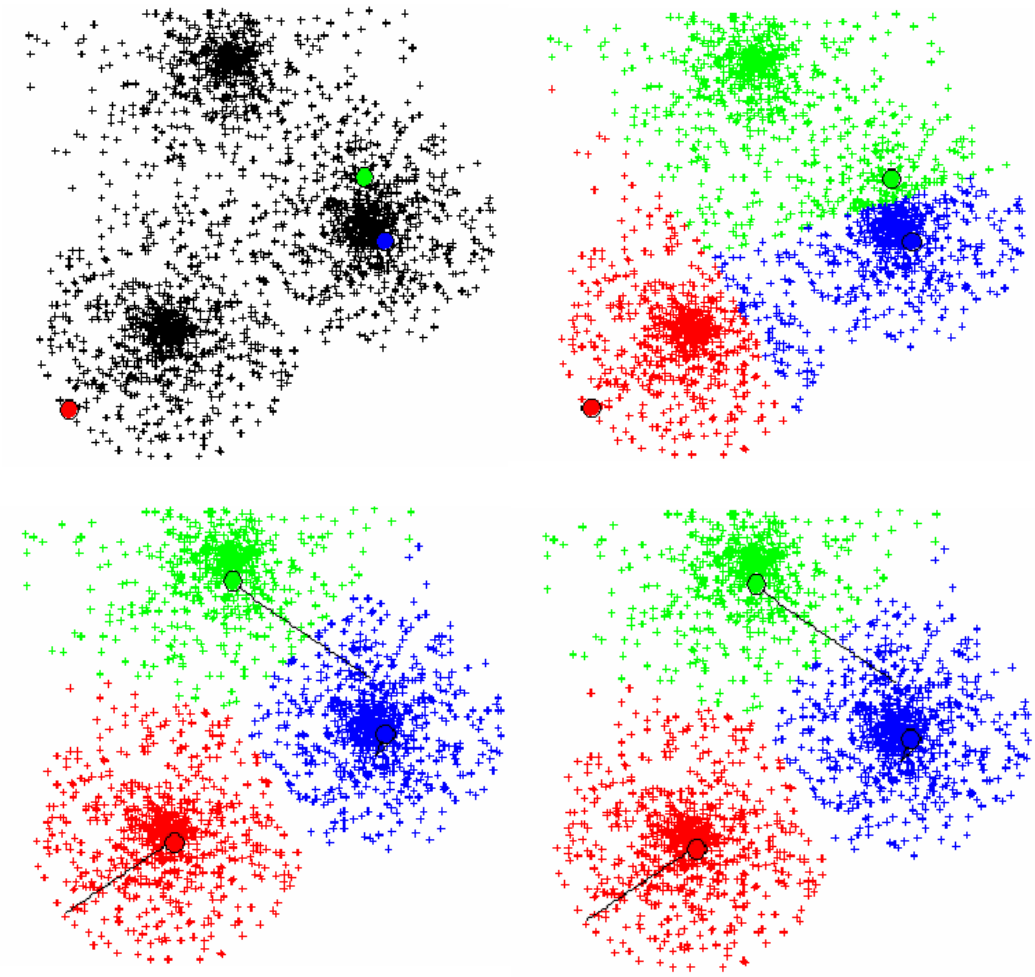

Figura 2: Etapas do algoritmo k-means – WEKA (Fonte: Santana, 2008).

4.4.3 Comparando os algoritmos Stars e K-means

Os dois algoritmos utilizados possuem grandes diferenças relativa a configuração e modo de processamento. Tais diferenças podem ser analisandas na Tabela 1.

Tabela 1: Comparação dos algoritmos Stars e K-means.

K-means	Stars
O usuário deve estabelecer a quantidade de grupos a serem gerados	Define automaticamente a quantidade de grupos.
Coloca os elementos aleatoriamente nas classes, calcula centróide de cada classe e verifica, para cada elemento, qual o centróide mais próximo.	Seleciona um 1 elemento e coloca todos os similares a ele no mesmo cluster. O grau de similaridade é estabelecido pelo usuário.
Refaz os agrupamentos iterativamente de modo que um elemento pode mudar de grupo ao longo do processo.	Depois de alocado em um grupo, um elemento não mais processado. Sendo assim, não garante que um elemento tenha sido alocado no grupo mais similar a ele.
Dependendo da quantidade de dados, demanda bastante tempo para processamento.	É rápido no processamento

4.5 A Base de Dados

A base explorada foi retirada do sistema de reclamações sobre produtos e serviços que armazena mensagens de usuários. Essa base recebe atualização diária, portanto, para a produção de material para esse estudo determinou-se que o conteúdo trabalhado deveria ser referente ao mês de agosto de 2009.

O conjunto de reclamações foi disponibilizado no formato .xls contendo duas colunas, sendo elas “Data” e “Descrição”. Para iniciar a preparação foi indispensável a separação dos documentos em arquivos textos individuais.

Segue ilustrado no Quadro 4 um exemplo de um documento que compõe a base.

Quadro 4 – Exemplo de documento que integra o Corpus

Efetuei um investimento em um equipamento da Dell Computadores (VOSTRO 1510), seduzido pelo discurso do televendas onde oferecem o céu e a terra para fechar a negociação, com menos de 1 mês de uso o equipamento apresentou problemas, fui muito bem atendido e o problema solucionado. Porem com aproximadamente 10 meses de uso o equipamento voltou a apresentar mais problemas, impossibilitando o seu funcionamento e uso, hoje fazem exatamente 90 dias que estou com o equipamento parado, minha garantia se expirou em 18/08/2009. Apos enorme persistência e inúmeras tentativas de troca de peças para solucionar o problema, consegui a troca do equipamento (30/07/2009), porem já faz 31 dias que estou aguardando a troca, exatamente hoje (31/08/2009) efetuei contato com a DELL e após exatamente 50 minutos ao telefone fui informado que eles nada podem fazer. Finalizando minha profunda insatisfação ainda fui seduzido para adquirir o pacote de serviços "COMPLETECARE", que segundo o vendedor me daria uma tranqüilidade a mais quando meu equipamento apresenta-se algum tipo de problema. Imaginem se não tivesse adquirido este pacote de serviços gentilmente oferecido pela DELL.

4.5.1 Escopo do Corpus

Um dos objetivos principais da mineração é a quantificação e caracterização de seu objeto de estudo. Com o levantamento dos números inerentes à base de dados textuais pode-se compreender a sua abrangência e iniciar a construção de deduções que antes estavam escondidas na forma de texto.

A descrição é iniciada observando a distribuição dos registros em relação ao tempo. A tabela 2 representa todos os documentos de agosto de 2009.

Tabela 2– Distribuição dos Documentos.

Data	Freqüência	%	Freqüência Acumulada	% Acumulado
1/1/2009	504	2,43	504	2,43
2/1/2009	1.084	5,22	1.588	7,65
3/1/2009	764	3,68	2.352	11,33
4/1/2009	591	2,85	2.943	14,18
5/1/2009	336	1,62	3.279	15,80
6/1/2009	1.396	6,73	4.675	22,53
7/1/2009	748	3,60	5.423	26,13
8/1/2009	838	4,04	6.261	30,17
9/1/2009	654	3,15	6.915	33,32
10/1/2009	1.033	4,98	7.948	38,30
11/1/2009	738	3,56	8.686	41,86
12/1/2009	467	2,25	9.153	44,11
13/1/2009	445	2,14	9.598	46,25

14/1/2009	358	1,73	9.956	47,98
15/1/2009	749	3,61	10.705	51,59
16/1/2009	734	3,54	11.439	55,12
17/1/2009	813	3,92	12.252	59,04
18/1/2009	460	2,22	12.712	61,26
19/1/2009	653	3,15	13.365	64,40
20/1/2009	679	3,27	14.044	67,68
21/1/2009	827	3,99	14.871	71,66
22/1/2009	278	1,34	15.149	73,00
23/1/2009	528	2,54	15.677	75,54
24/1/2009	799	3,85	16.476	79,39
25/1/2009	859	4,14	17.335	83,53
26/1/2009	596	2,87	17.931	86,41
27/1/2009	697	3,36	18.628	89,76
28/1/2009	738	3,56	19.366	93,32
29/1/2009	487	2,35	19.853	95,67
30/1/2009	330	1,59	20.183	97,26
31/1/2009	569	2,74	20.752	100,00

Observando a Tabela 2, verifica-se a existência de 20.752 registros. A Figura 3 mostra um gráfico que ilustra a distribuição de registro de forma mais clara.

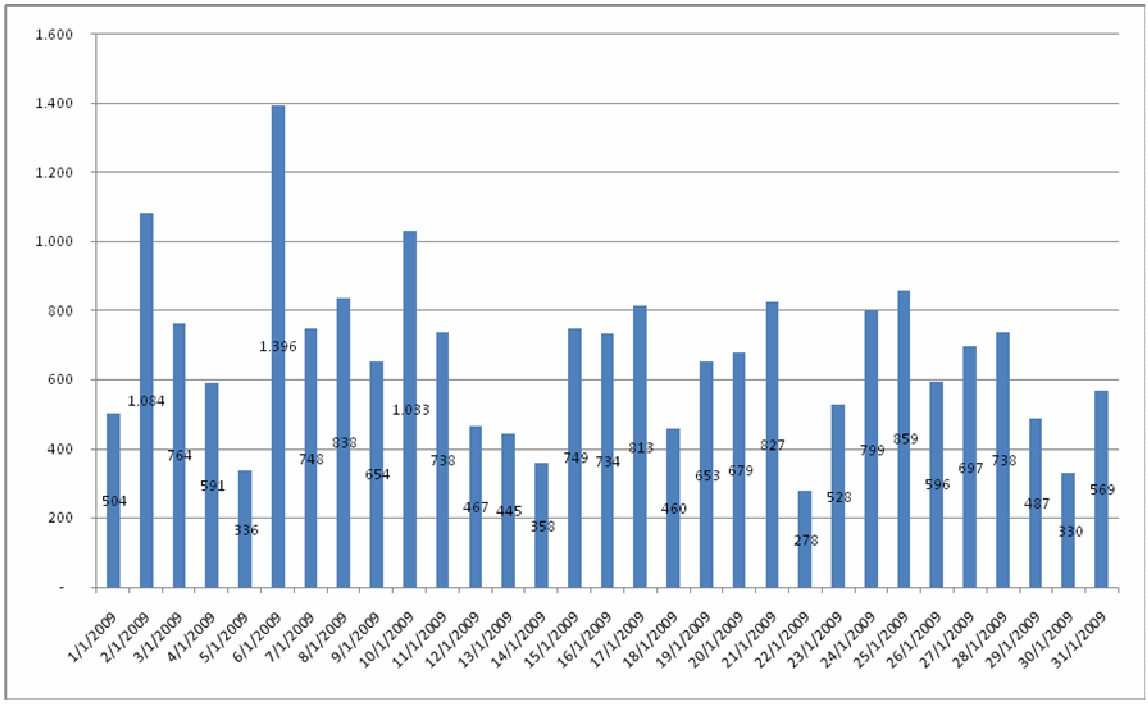

Figura 3 – Gráfico de Distribuição dos Documentos.

4.6 Fluxo Operacional

Na base de reclamações estão contidos os documentos necessários para o estudo. Os arquivos textuais vão alimentar o processo de mineração de texto que determinarão agrupamentos cujo objetivo é reunir os documentos similares.

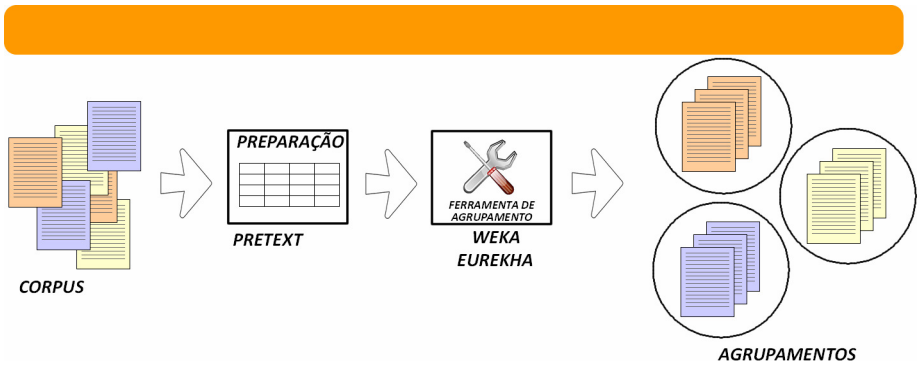

Figura 4: Fluxo Operacional da Pesquisa.

4.7 O Corpus e a Preparação de Dados

Realizada a coleta e seleção dos documentos, um passo demorado e delicado a se fazer é o pré-processamento dos documentos. O descuido ou a má realização de alguma das tarefas que compõem esta etapa pode comprometer todo o projeto de mineração. Estas tarefas englobam desde a remoção de registros duplicados, até a supressão de caracteres insignificantes como se poderá observar a diante.

A primeira tarefa para se apurar um texto sem estrutura é identificar suas características. Para isso é necessário dividir documento em *tokens* ou termos. Esse processo é trivial para uma pessoa que possua conhecimento da estrutura de linguagem. Porém, para um programa de computador, essa tarefa pode ser um pouco mais complicada. Para realizar esse processo, é necessário remover alguns caracteres indesejados, tais como sinais de pontuação, separação silábica, marcações especiais e números, os quais, isoladamente, fornecem pouca informação. Finalmente a extração dos termos é executada utilizando normalmente o espaço em branco entre as palavras como indicador para dividir o texto em termos.

4.7.1 Stemming

Grande parte do trabalho de pré-processamento dos dados foca a diminuição no escopo do *corpus*, isto é, a redução do número de termos que alimentam a ferramenta de mineração de textos. A ferramenta Pretext possui recursos que viabilizam esta tarefa. A tarefa de *stemming* resume-se em uma normalização lingüística, na qual as formas diferentes de um termo são reduzidas a uma forma comum chamada *stem*. O resultado da aplicação de algoritmos de *stemming* incide na remoção de prefixos e sufixos de um termo. Por exemplo, os *tokens*: `informar`, `informação`, `informações`, `informando`, `informou`, `informado` e `informe`, podem ser transformados parra um mesmo *stem* `inform` (Figura 5).

Os algoritmos de *stemming* possuem forte dependência do idioma no qual os documentos estão escritos. Para a língua portuguesa, existem algoritmos de *stemming* que foram adaptados do algoritmo de Porter. Nesses algoritmos, os sufixos dos *tokens*, com um comprimento mínimo estabelecido, são eliminados considerando se algumas regras pré-estabelecidas. Caso não seja possível eliminar nenhum sufixo de acordo com essas regras, são analisadas as terminações verbais da palavra. Essa é a principal diferença entre o algoritmo de *stemming* para palavras em inglês e para palavras em português ou espanhol, por exemplo. Enquanto na língua inglesa a conjunção dos verbos é quase inexistente para verbos regulares, pois usualmente acrescenta-se a letra s no final do verbo no presente na terceira pessoa do singular, as linguagens provenientes do latim apresentam formas verbais altamente conjugadas em sete tempos, que contem seis terminações diferentes cada tempo (SOARES *et al,* 2008).

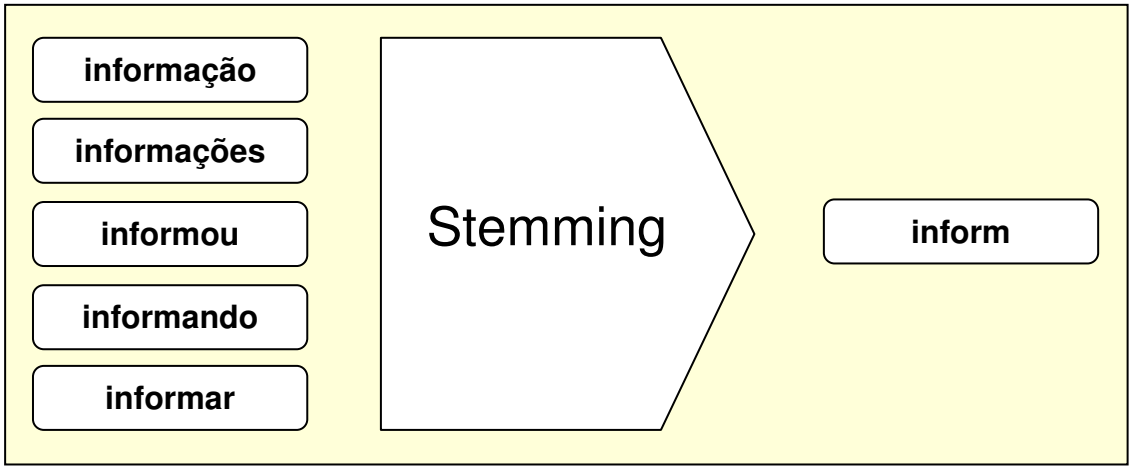

Figura 5 – Processo de Stemming nos Termos

É pouco provável que o algoritmo de *stemming* retorne o mesmo *stem* para todos os *tokens* que tenham a mesma origem ou radical morfológico, pois a maioria dos algoritmos de *stemming* ignoram o significado dos termos, introduzindo alguns erros. *Tokens* com significados diferentes podem ser reduzidos a um mesmo *stem*, nesse caso ocorre um erro de *over-stemming*. Por exemplo, os *tokens* `barato` e `barata` podem ser reduzidos a um mesmo stem `barat`. Já *tokens* com significados similares quando reduzidos a *stem* diferentes conduzem ao erro de *under-stemming*. Por exemplo, os *tokens* `viagem` e `viajar` podem ser reduzidos aos *stems* `viag` e `viaj`, respectivamente.

Outro tipo de erro, denominado de *mis-stemming*, consiste em retirar o sufixo de um *token* quando na verdade essa redução não é necessária. Por exemplo, o *token* `lápis` poderia ser reduzido ao *stem* `lapi`, dependendo de como o plural das palavras é tratado pelo algoritmo de *stemming*.

Já foi analisado que à medida que o algoritmo se torna mais específico na tentativa de reduzir a quantidade de *tokens* distintos para palavras com um mesmo radical, a eficiência do algoritmo degrada (PORTER, 1980).

Durante o processo de *stemming*, pode-se observar que houve *stem* que foi criado da redução de até 16 *tokens* de palavras semelhantes. Conforme Quadro 5:

Quadro 5: Exemplo de Stem realizado no corpus.

```
cancel : 110(54/300)
       cancelavam:1
       cancelasse:1
       cancelaram:1
       cancelarei:1
       cancelarem:1
       cancelados:2
       cancelaremos:2
       cancelo:3
       cancelada:3
       cancelamos:3
       cancelando:4
```

```
cancelado:6
cancelei:7
cancelou:8
cancelar:25
cancelamento:42
```

O Quadro 5 indica que os 16 termos distintos entre si, aparecem em 54 dos 300 documentos. Indica também que somadas o total de suas aparições resultam em 110 vezes. Todos esses termos foram reduzindo ao *stem* "`cancel`".

4.7.2 Termos Freqüentes

Seguindo com o pré-processamento, sabe-se que todos os documentos textuais, independente da língua em que estão escritos, várias palavras são muito comuns e não são significativas para o algoritmo de aprendizado. Essas palavras incluem pronomes, artigos, preposições, advérbios, conjunções, entre outras. Essas palavras são denominadas *stowords*. Assim é gerada uma *stoplist*, na qual são armazenadas inúmeras *stopwords* com a finalidade de serem desconsideradas durante o processo de mineração, diminuindo assim o número de *tokens* utilizados para representar os documentos.

Prosseguindo com a redução do escopo do corpus, foi necessária então a criação de lista de *stopwords*. A lista criada foi criada no padrão XML e possui 680 *stopwords.* Este procedimento representou grande redução na dimensionalidade no *corpus* trabalhado.

Quadro 6: Exemplo da Lista de Stopwords criada.

```
<?xml version="1.0" encoding="utf-8"?>
<stopfile>
   <stopword>a</stopword>
   <stopword>abaixo</stopword>
   <stopword>acaso</stopword>
   <stopword>acerca</stopword>
   <stopword>acima</stopword>
   <stopword>acola</stopword>
   <stopword>ademais</stopword>
   <stopword>adentro</stopword>
   <stopword>adiante</stopword>
   <stopword>afinal</stopword>
   <stopword>afora</stopword>
```

```
   <stopword>agora</stopword>
   <stopword>agorinha</stopword>
   <stopword>ai</stopword>
</stopfile>
```

Termos que aparecem em todos os documentos não possuem muito valor informativo para a tarefa de criação de agrupamentos, pois não são capazes de discriminar um documento de outro. Com o auxílio da lista ilustrada na Quadro 6, grande parte destes termos foram eliminados.

4.7.3 Eliminando Duplicação

No decorrer da etapa de pré-processamento, com a utilização da ferramenta Excel, antes de separar os registros em arquivos textos individuais, foi realizado o rastreamento de documentos duplicados.

Notou-se que existiam 978 registros em duplicidade, porém, excluí-los de imediato poderia retirar informação importante do *corpus*. Diante disso, foi realizada a verificação registro a registro e constatou-se que todas as supostas duplicidades eram positivas, com isso foi realizada a eliminação das duplicidades. Segue na Tabela 3 exemplo de registros duplicados:

Tabela 3 – Exemplo de Registros Duplicados

De acordo com o termo de uso aceito por todos os cadastrados no Reclame Aqui, o USUÁRIO não poderá incluir comentários ilícitos no site do Reclame Aqui!, De forma a atribuir a alguém a prática de crime, imputar a alguém fato ofensivo à sua reputação, e, ofender alguém atentando contra sua dignidade ou decoro. Caso o conteúdo de sua reclamação esteja em desacordo com a cláusula citada acima, o Reclame Aqui poderá excluir sua reclamação. CLIQUE AQUI e escreva sua reclamação.
De acordo com o termo de uso aceito por todos os cadastrados no Reclame Aqui, o USUÁRIO não poderá incluir comentários ilícitos no site do Reclame Aqui!, De forma a atribuir a alguém a prática de crime, imputar a alguém fato ofensivo à sua reputação, e, ofender alguém atentando contra sua dignidade ou decoro. Caso o conteúdo de sua reclamação esteja em desacordo com a cláusula citada acima, o Reclame Aqui poderá excluir sua reclamação. CLIQUE AQUI e escreva sua reclamação.

Os registros apresentados na Tabela 3 indicam uma duplicação. Em média cada duplicação ocorreu na ordem de uma vez, entretanto, no caso de

registro ilustrado na Tabela 3 a duplicação ocorreu em uma ordem de 645 vezes. A exclusão destes registros representou uma grande melhora na qualidade do *corpus*.

4.7.4 Representação Quantitativa do Texto

Uma parte essencial da mineração de textos está na adequação do texto ao formato que pode ser reconhecido por algoritmos de mineração, isto é, na transformação do texto em números atentando em não perder a informação nele codificada.

Em bases de dados estruturadas, os dados são representados em formato de tabela atributo-valor. Assim, cada linha refere-se a um documento e cada coluna refere-se a um atributo especifico.

Tabela 4 – Representação de Dados.

VEICULO	CAPACIDADE	CONSUMO	PREÇO
Gol	5	12	22
Fusca	5	8	3
...	...	...	...
Kombi	12	7	11

Pode-se observar na Tabela 4 que as linhas se referem aos veículos e as colunas, aos atributos relacionados aos veículos. Sendo assim, um veículo é composto pela soma de seus atributos. Essa é a forma convencional de se representar os dados e utilizá-los para procedimentos de descoberta de conhecimento em dados.

Em um documento, pode-se representar os atributos através de seus termos de maneira que possamos dispô-los em tabelas nos mesmos moldes. Assim, cada documento de uma coleção é representado pelas linhas e cada coluna refere-se ao termo contido no texto correspondente.

Como exemplo, tem-se uma coleção de documentos e cada documento com seu teor textual:

Doc1 – “A informação é importante”;

Doc2 – “Necessitamos de tratamento para a informação”;

Doc3 – “Através de tratamento é obtido informação importante”.

Sendo assim, pode ser representado conforme a tabela 5, como se segue:

Tabela 5 – Representação do Corpus.

Docs		Termos									
	a	**informação**	**é**	**importante**	**necessitamos**	**de**	**tratamento**	**para**	**através**	**que**	**obtido**
Doc1	S	S	S	S	N	N	N	N	N	N	N
Doc2	S	S	N	N	S	S	S	S	N	N	N
Doc3	N	S	S	S	N	S	S	N	S	S	S

Na da Tabela 5, é ilustrada uma representação binária informando a presença “S” do termo ou não “N”. De maneira semelhante à Tabela 3, um documento é representado pela soma de todos os seus atributos alistados.

Considerando ***n*** documentos, a quantidade de documentos na coleção representados por $D = \{D_1, D_2, ... D_n\}$, e ***m*** termos, ou atributos, presentes no corpus representados por $T = \{T_1, T_2, ... T_n\}$. Cada documento ***D*** é representado por ***m*** termos contidos no documento. Cada termo pode ser uma palavra simples ou composta. Então, A_{nm} representa a influência do atributo m no documento n que pode estar representada pela indicação da presença do termo, pela freqüência do termo em relação ao documento. Portanto, qualquer *corpus* pode ser representado conforme Tabela 6.

Tabela 6 – Representação generalizada do Corpus.

Documentos	Termo			
	T1	**T2**	**T3**	**T4**
D1	a11	a12	**...**	A1m
D2	a21	a22	**...**	A2m
...	**...**	**...**	**...**	**...**
Dn	an1	an2	**...**	anm

Desse modo, as táticas para quantificar os termos na tabela são variadas. Em algumas situações, pode-se levar em consideração a simples existência do termo, em outros, a freqüência do termo em relação ao documento. A escolha da representação depende do tipo de aplicação que se deseja executar.

Representação Binária – leva em consideração a simples existência do termo. Se o termo ocorrer no documento, o valor de a_{ij} é 1, caso contrário, a_{ij} recebe recebe valor 0.

Reescrevendo a Tabela 6 temos a seguinte representação da Tabela 7:

Tabela 7 – Representação em Código Binário.

Docs	Termos										
	a	**informação**	**é**	**importante**	**necessitamos**	**de**	**tratamento**	**para**	**através**	**que**	**obtido**
Doc1	1	1	1	1	0	0	0	0	0	0	0
Doc2	1	1	0	0	1	1	1	1	0	0	0
Doc3	0	1	1	1	0	1	1	0	1	1	1

Esse tipo de representação certifica somente a existência do termo, não considerando a quantidade de vezes em que ele aparece. Para algumas aplicações, mais simples, esta representação é satisfatória.

Representação por Freqüência - considera-se a contagem de vezes que o termo ocorre. Essa medida é repetidamente apresentada com tf, do inglês *"term frequency".* Esse tipo de representação possibilita a idéia da importância de um termo conforme a sua presença.

O termo a_{ij} é atribuído do valor de tf (tj, di) que é a freqüência do termo tj no documento di.

Sua representação se compõe conforme disposta na Tabela 8:

Tabela 8 – Representação do Corpus em Freqüência.

Docs	Termos										
	a	**informação**	**é**	**importante**	**necessitamos**	**de**	**tratamento**	**para**	**através**	**que**	**obtido**
Doc1	1	1	1	1	0	0	0	0	0	0	0
Doc2	1	1	0	0	1	1	1	1	0	0	0
Doc3	0	1	1	1	0	1	1	0	1	2	1

4.7.5 Explorando o Texto

Continuou-se a procura de problemas na base de documentos.

A princípio a coleção de documentos possui 3.930.629 termos distribuídos em 19.774 documentos.

Ao realizar a contagem dos registros em branco, identificou-se a presença de termos com conteúdos muito curtos e que poderiam indicar um erro. Segue na Tabela 9 todos os documentos com menos de 10 termos:

Tabela 9 – Documentos com menos de 10 termos.

Cont.	***Reclamação***	***Qdade Termos***
1º	serhsdfhbsnserhrtrts	1
2º	Olá	1
3º	Negociar	1
4º	Boa tarde..	2
5º	teste 123	2
6º	um mes	2
7º	queria vender minha divida	4
8º	quero cancelar o superzig.	4
9º	NÃO TENHO NENHUMA RECLAMAÇÃO.	4
10º	17/6/09 Cancelamento de pedido	4
11º	quero negociar minhas dividas .	4
12º	Não querem cancelar nossa conta!	5
13º	seu incompetentes,cade minha comoda	5
14º	quero cancelar ar mensagem do 48022	6
15º	quero ver minhas faturas pela internet	6
16º	Comprei um chuveiro e ele está ruim.	7
17º	NAO QUERO MAIS ESSE SERVIÇO POR FAVOR	7

18º	o chip bloqueado e o telef. 65 81375325	7
19º	eu não consigo cancela o clube movilisto	7
20º	Bom dia, ainda não recebi o cabo USB.	8

Fazendo uma analise na Tabela 9, pode-se verificar que do 1º ao 6º registro, tratam-se de erros de digitação ou de mensagens que, devido a sua insignificância, não somam nenhum valor para o corpus. Com isso, decidiu-se excluí-las, restando agora 19.768 documentos.

4.8 Amostragem

Diante do grande número de documentos, considerando a restrição de desempenho do equipamento utilizado no trabalho, não foi possível processar todo o volume de documentos pertencentes ao *corpus*. Porém este problema pode ser vencido com o uso de amostra. Dessa maneira foi extraída uma amostragem contendo 300 documentos do total do corpus. O critério escolhido foi à amostragem estratificada afim de não perder o valor representativo da coleção de dados.

A amostragem estratificada resume-se em separar a população geral em estratos, que podem ser entendidos como sub-populações ou subconjuntos. Deve-se cuidar para que o comportamento dentro de cada estrato seja razoavelmente homogêneo. Em tais situações, se o sorteio dos elementos for executado sem considerar a existência dos estratos, pode acontecer de alguns estrados não serem representados na amostra, o que irá influenciar no resultado devido à ocorrência de estratos mais favorecidos pelo sorteio.

Na amostragem da base em questão, foi definido que cada dia do mês representaria um estrato, resultando num total de 31 estratos, com uma média da população geral, aprontou-se uma amostra de 300 documentos, que devido aos cuidados na elaboração da amostra, minimizará a perca de valores característicos da população geral.

5 Resultados

De posse dos dados e ferramentas adequadas para a realização do projeto, foram executadas as tarefas descritas na metodologia e que se apresentam no decorres deste capítulo.

5.1 Ponderação dos Termos

Visto que o trabalho de depuração da base se achava em um grau satisfatório, efetuou-se o agrupamento da coleção com o propósito de alocar os documentos similares em grupos. A finalidade é elevar ao máximo a diferença entre os grupos e tornar mínimo a diferença internamente.

A criação desses grupos depende da seleção do critério de transformação em números. No projeto, foi utilizada a ponderação *Term Frequency* (TF). Analisando-se a variância interna, observa-se que a medida TF obteve melhor resultado, permitindo a criação de uma quantidade maior de agrupamentos, resultando em grupos mais homogêneos.

5.2 Aplicação do algoritmo k-means com o WEKA

A aplicação do algoritmo *k-means* com WEKA gerou em um resultado satisfatório.

Uma vez que este algoritmo necessita que o usuário especifique o número de *cluster* a ser gerado e número de sementes (*seed*), o melhor resultado adquirido, depois de realizar vários testes, foi com o número de *cluster* igual a “7” e semente igual a “10”. Com essa configuração, no final da execução do algoritmo conseguiu-se identificar sete grupos no *corpus*, sendo um descartado por não fornecer significância, restando 6 grupos. Essa quantidade de *cluster* foi a que melhor representou a distribuição dos documentos nos respectivos grupos. Isto pode ser conferido nos grupos conforme a Figura 6.

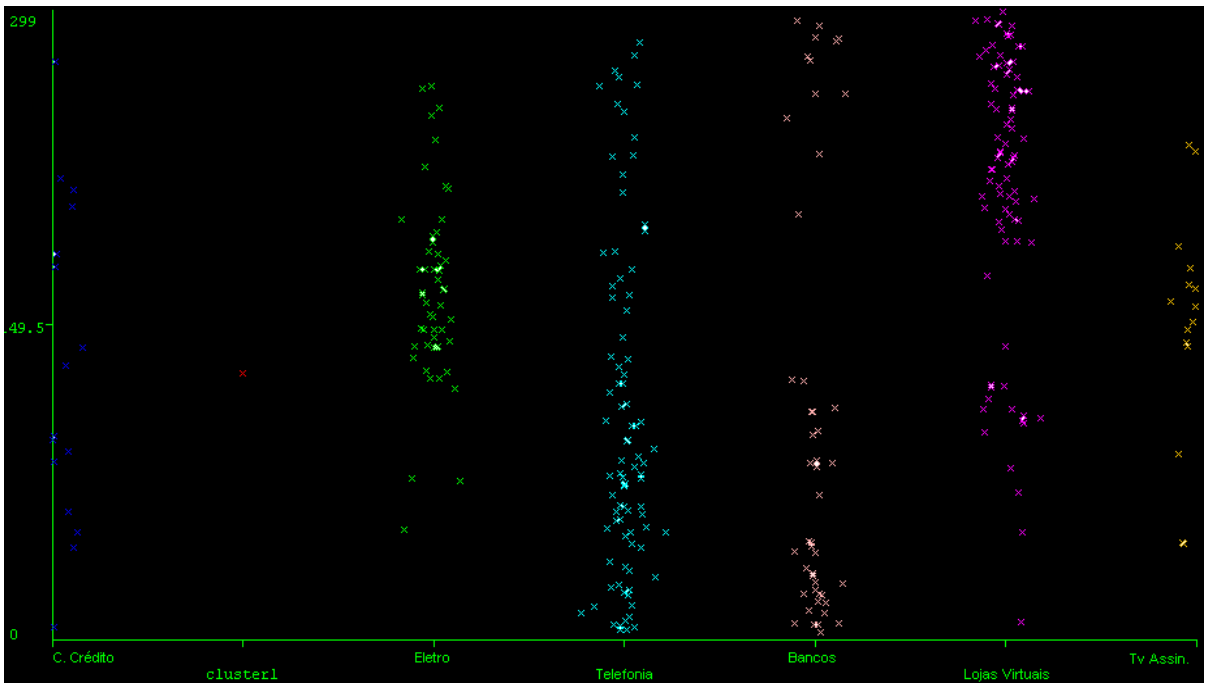

Figura 6: Grupos gerados pelo algoritmo k-means no WEKA..

Na Figura 7 segue um gráfico da distribuição dos grupos gerados.

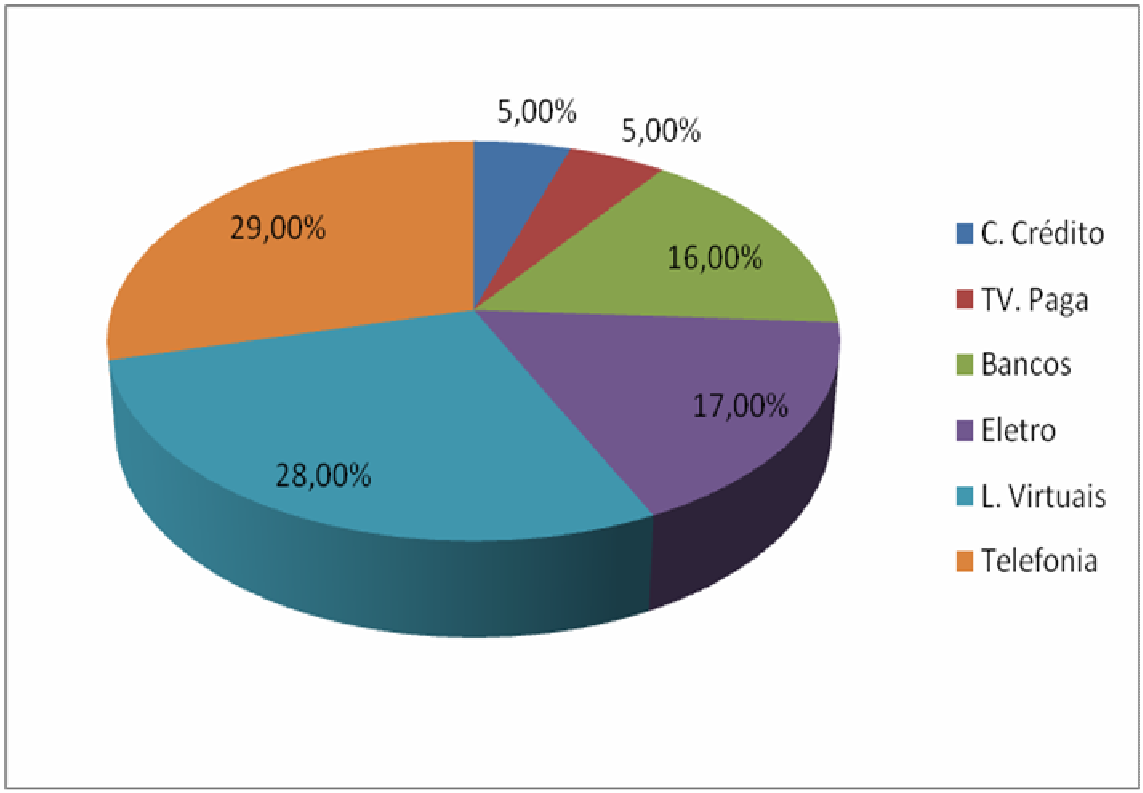

Figura 7: Distribuição de grupos gerados pelo WEKA.

Analisando o gráfico da Figura 7, nota-se que as empresas de Telefonia lideram o ranking como segmento mais reclamado, seguido das Lojas

Virtuais (e-commerce), Eletro (Eletroeletrônicos, Eletrodomésticos e equipamentos), Bancos e Financeiras, Cartão de Crédito e TV por assinatura.

A identificação dos assuntos pertencentes a cada *cluster* foi realizada através dos atributos pertencentes em cada agrupamento. Por exemplo, o *cluster* 04 (telefonia) foi formado pelos atributos tel, recarg, celul, tarif, entre outros. Avaliando estes atributos pode-se definir a qual segmento tal agrupamento se referia.

5.3 Aplicação do algoritmo Star com o Eurekha

Para melhor aproveitar os dados, aplicou-se também a mineração através da ferramenta Eurekha utilizando o algoritmo *Star*.

No final da aplicação deste algoritmo, notou-se que o mesmo não mostrou eficácia na obtenção de resultados, conforme Figura 8.

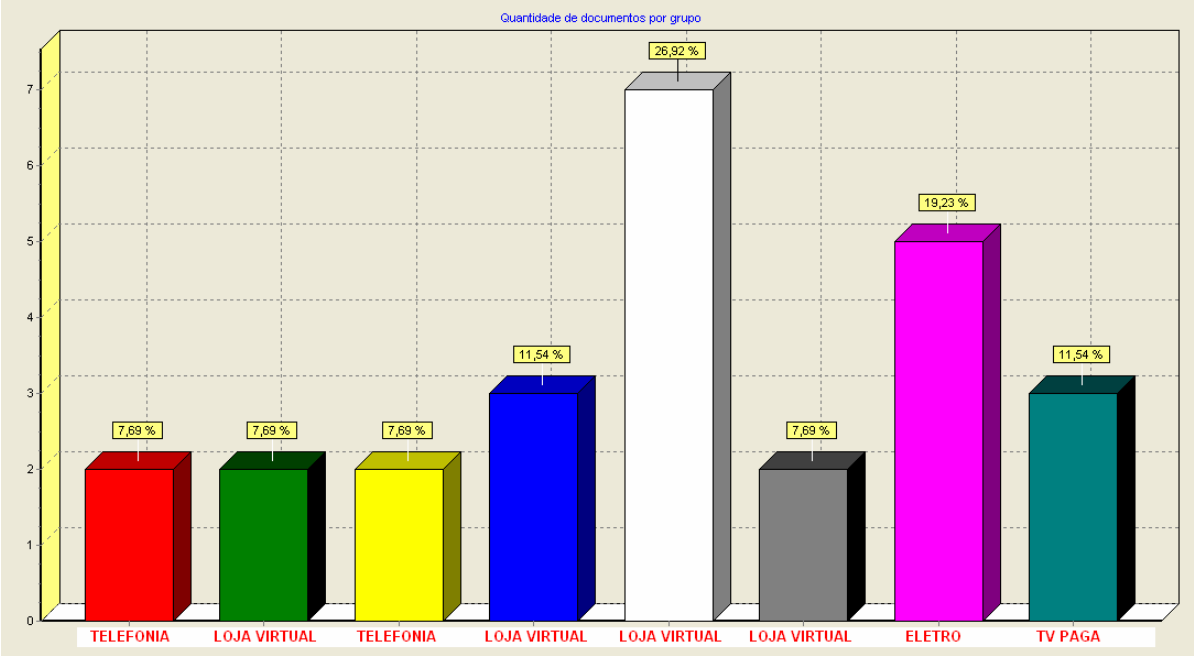

Figura 8: Resultado do agrupamento através do algoritmo Star.

Fazendo uma análise no resultado ilustrado na Figura 8, pode-se verificar que o algoritmo não conseguiu identificar os grupos Cartão de Crédito e Bancos. Este resultado é justificado pelo motivo de ter sido aplicado em uma base que possui vários conceitos distintos, que abrange diversas situações e muitas vezes os atributos de cada documento não é suficiente para classificá-lo em um grupo correto.

No caso ilustrado na Figura 8, nota-se que o conceito “Loja Virtual” é repetido quatro vezes. Isso se dá pelo fato de ser um conceito muito amplo, onde inclui transações como pagamentos, cheques, cartões (que em alguns casos deveria ser alocado no grupo “Bancos” ou “Cartão de Crédito). Diante desse problema, o algoritmo não foi capaz de diferenciar com qualidade os documentos e não identificou todos os conceitos presentes no corpus.

5.4 Comparações de resultados

Com o fim dos trabalhos realizados com os algoritmos escolhidos e com posse de dados percentual das reclamações fornecidas pelo especialista do sistema de reclamações (Tabela 10), pode-se realizar uma comparação entre o agrupamento original e os agrupamentos gerados através das ferramentas de mineração.

Tabela 10: Ranking de reclamações fornecidas pelo especialista do Sistema de Reclamações.

SEGMENTOS MAIS RECLAMADOS		
Posição	**Reclamada**	**Reclamações**
1°	**Telefonia Fixa, Móv...**	84506
2°	**Lojas Virtuais - Co...**	71015
3°	**Eletroeletrônicos, ...**	31605
4°	**Bancos e Financeiras**	19194
5°	**Tv ,tv Por Assinatu...**	15368
6°	**Cartões de Crédito**	13916

Na Tabela 11, foram colocados os três grupos para se comparar os resultados.

Tabela 11: Comparação dos resultados.

	Weka/k-means	Dados Oficiais	Eurekha/Star
C. Crédito	5,00%	4,80%	-
TV. Paga	5,00%	5,35%	11,54%
Bancos	16,00%	6%	-
Eletro	17,00%	9,88%	19,23%
L. Virtuais	28,00%	22,19%	53,84%
Telefonia	29,00%	26,41%	15,38%

Analisando a Tabela 11, pode concluir que diferente do algoritmo *Star*, o *k-means* obteve um resultado satisfatório ao ser comparado com dados oficiais. É visível que existem certas diferentes que variam entre os grupos (grupo "Eletro" por exemplo). Porém, vale lembrar que foi abordada uma base de reclamação sobre produtos e serviços, que abrange assuntos diversos e é alimentada por usuários distintos. Também se deve considerar que a base inicial possuía 20.752 documentos, devido à capacidade de processamento das ferramentas utilizadas, foi retirada uma amostragem de 300 documentos. Apesar de utilizar uma boa técnica para realizar a amostragem, a perda de qualidade na representação do *corpus* é inevitável. Mas mesmo diante de problemas o resultado obtido ainda foi aceitável.

6 Conclusão

O trabalho objetivou o tratamento automático da informação textual com a menor invenção humana possível.

Nota-se que a informação textual ainda necessita de profissionais e ferramentas, com alto poder de processamento, capaz de manipular com eficiência grande quantidade de dados sem nenhuma estrutura.

Para chegar aos objetivos propostos fez-se o uso do processo de DCT que abrange desde a escolha da base até aplicação a eficaz das ferramentas para retirada de conhecimento dos dados.

A base de dados que foi disponibilizada para mineração não estava preparada e apresentou diversos problemas. Foi preciso a utilização de programa para separar e transformar os dados que vieram em planilhas .xls para documentos .txt individuais. Para a tarefa de transformação e separação da base em documentos textuais individuais, utilizou-se a ferramenta Intext. Antes dessa separação foi necessário tratar a duplicação de registros. Esse problema foi resolvido com o uso do Excel, identificando e eliminando os registros duplicados, porém, foi um processo que demandou certo tempo.

Erros de ortografia e os erros de pontuação, se não tratados, comprometem o resultado final e, portanto, esse tratamento é um ponto fundamental no decorrer do projeto. Para esse trabalho, a intervenção humana foi indispensável.

Durante o passo de descrição da base, pode-se visualizar o valor de negócio que a informação do consumidor, na forma de reclamação, promove. Empresas podem utilizar deste canal direto com o cliente e através de suas informações melhorar seus processos, produtos e serviços. Com o atendimento da reclamação por parte da empresa, o cliente se sente estimado e firma uma relação duradoura que resulta no continuo consumo dos serviços e produtos oferecido pela empresa.

Considera-se que o objetivo dessa pesquisa foi alcançado, pois foi proposto estudar e aplicar técnicas de mineração de textos, extrair conhecimentos da base de reclamações sobre produtos e serviços, extrair conhecimento da base de dados e criar categorias de agrupamentos de forma automática para os documentos.

Tabela 12: Cumprimento dos objetivos.

OBJETIVOS ESPECIFICOS	***RESULTADOS***
Estudar as técnicas de mineração.	Estudo e descrição das técnicas e meio de aplicação de mineração de textos em base não estruturada.
Aplicar as tarefas de mineração em uma base de dados sobre produtos e serviços	Realização de coleta, pré-processamento, extração de informação e avaliação do conhecimento sobre os textos.
Extrair conhecimentos da base de dados	Estatísticas descritivas sobre o Texto, número de termos, apontamento de duplicidade nos documentos, além de problemas destacados na pesquisa.
Categorizar os documentos automaticamente	Criação de agrupamentos utilizando dois algoritmos de categorização.

Analisando a Tabela 12, conclui-se que a metodologia sugerida foi eficaz e aplicável em realizar a transformação de dados textuais sem estrutura em informações organizadas, permitindo a extração de conhecimentos, que entes eram implícitos nos mesmo.

7 Considerações finais e trabalhos futuros

No processo de mineração de textos, todas as etapas, desde a coleta de documentos até a extração de conhecimentos, são de extrema importância e demandam grande atenção dispensada em cada uma delas. A etapa mais trabalhosa, sem dúvidas é a de pré-processamento, isso pelo fato de agregar diversas tarefas que refletem diretamente nos resultados finais. A cada etapa depende da boa realização de etapa anterior. A boa realização de cada uma delas incide fortemente no sucesso do trabalho com um todo.

Ao se realizar este trabalho, despontaram novas idéias para a continuação futura do mesmo. Como o algoritmo *k-means* considera cada atributo como uma dimensão, o que eleva o custo computacional do processamento. Com isso, pode-se vir a aplicar um algoritmo de análise fatorial ou similar sobre os dados, a fim de diminuir a quantidade de dimensões, outra alternativa, seria a utilização de *feature selection* com a finalidade de reduzir a quantidade de palavras.

Fica também a proposta da implementação dos algoritmos estudados em uma ferramenta própria, focando a análise da base de reclamações, de maneira a poder utilizar todos os documentos fornecidos.

9 Referências

BARLOW, J. & MOLLER, C. **Reclamação de cliente? Não tem melhor presente**. São Paulo: Futura, 1996.

MARTINS, Claudia Aparecida. **Uma Abordagem para pré-processamento de dados textuais em algoritmos de aprendizado**. 2003. 31 p. Instituto de Ciências Matemáticas e de Computação – ICMC – USP, São Paulo.

MOONEY, Raymond J., BUNESCU, Razvan. **Mining Knowledge from Text Using.** 2005. Disponível em: <http://www.sigkdd.org/explorations/issues/7-1-2005-06/2-Mooney.pdf>. Acesso em: 11 de maio de 2009.

MOONEY, Raymond J., BUNESCU, Razvan. **Statistical Relational Learning for Natural Language Information Extraction.** 2005. Disponível em: <http://www.cs.utexas.edu/users/ml/papers/srl-submitted-05.pdf>. Acesso em: 11 de maio de 2009.

PORTER, M. F. **An algorithm for suffix stripping. Program**. 1980. Disponível *em:* <http://tartarus.org/~martin/PorterStemmer/def.txt>. Acessado em 28 de agosto de 2009.

RAJMAN, M.; BESANÇON, R. **Text Mining:** Natural Language techniques and Text Mining applications, Chapman & Hall, 1997. Disponível em: <http://eprints.kfupm.edu.sa/68734/>. Acesso em: 02 de julho de 2009.

REZENDE, Solange Oliveira. **Sistemas Inteligentes**: Fundamentos e Aplicações 2005. São Paulo: Manolle. p 337-370.

RINO, Lucia Helena Machado; PARDO, Tiago Alexandre Salgueiro. 2003. **Sumarização Automática de Textos:** Principais Características e metodologias. Disponível em: <http://www.icmc.usp.br/~taspardo/JAIA2003-RinoPardo.pdf>. Acesso em: 14 de maio de 2009.

SANTANA, Rodolfo Charamba. **Uma aplicação de CBIR à análise de imagens médicas de imuno-histoquímica utilizando Morfologia Matemática e espectro de padrões.** 2008. Disponível em: <http://dsc.upe.br/~tcc/20082/Rodolfo%20Charamba.pdf>. Acesso em: 04 de novembro de 2009.

SANTOS, Rafael. **Weka na Munheca.** Um guia para uso do Weka em *scripts* e integração com aplicações em Java. 2005. Disponível em: <http://www.lac.inpe.br/~rafael.santos/Docs/CAP359/2005/weka.pdf>. Acesso em: 02 de novembro de 2009.

SCHIESSL, José Marcelo. 2007. **Descoberta de Conhecimento em Texto:** Aplicada a um sistema de atendimento ao consumidor. Disponível em: <http://bdtd.bce.unb.br/tedesimplificado/tde_busca/arquivo.php?codArquivo=1538>. Acesso em 03 de junho de 2009.

SOARES, Matheus V. B., PRATI, Ronaldo C., MONARD, Maria C. **PreTexT: A Reestruturação da Ferramenta de Pré-Processamento de Textos**. 2008. Disponível em: <http://www.icmc.usp.br/~caneca/pretext.htm>. Acessado em: 28 de agosto de 2009.

TAN, A Hwee. **Text Mining:** The state of the art and the challenges. 1999. Disponível em: <http://www3.ntu.edu.sg/home/asahtan/Papers/tm_pakdd99.pdf>. Acesso em: 14 de maio de 2009.

WIVES, Leandro K. **Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "Clustering".** 1999. Disponível em: <http://www.leandro.wives.nom.br/pt-br/publicacoes/dissertacao.pdf>. Acesso em: 02 de novembro de 2009.