



**COPPE/UFRJ**

## CLASSIFICAÇÃO DE TEXTO PARA MELHORIA NO ATENDIMENTO DE HELP DESK

João Alquéres de Avellar Menezes

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro, RJ – Brasil

Outubro de 2009

CLASSIFICAÇÃO DE TEXTO PARA MELHORIA NO ATENDIMENTO DE  
HELP DESK

João Alquéres de Avellar Menezes

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO 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 MESTRE EM  
CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO

Aprovada por :

---

Prof. Geraldo Bonorino Xexéo, D. Sc.

---

Prof. Geraldo Zimbrão da Silva, D. Sc.

---

Prof. Jorge Lopes de Souza Leão, D. Ing

Rio de Janeiro, RJ – Brasil

Outubro de 2009

Menezes, João Alquéres de Avellar

Classificação de textos para melhoria do atendimento de Help  
Desk / João Alquéres de Avellar Menezes - Rio de Janeiro:  
UFRJ/COPPE, 2009.

X, 64 p. il.; 29,7cm

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de  
Engenharia de Sistemas e Computação, 2009.

Referências Bibliográficas: p. 58-64.

1. Aprendizado de máquina I. Xexéo, Geraldo Bonorino. II.  
Universidade Federal do Rio de Janeiro, COPPE, Programa de  
Engenharia de Sistemas e Computação. III. Título.

## Agradecimentos

A Deus, que torna todas as coisas possíveis.

Aos meus pais, Julio e Lucia, por todas as oportunidades que me deram.

A minha avó, Maria Cecília, por todo o apoio nesses anos.

A minha namorada, Carina, que me acompanhou e incentivou durante essa jornada, pela valiosíssima ajuda, esclarecimentos, revisões e principalmente apoio.

Ao meu orientador, o professor Xexéo, com quem os anos de convivência só aumentaram meu respeito e admiração, pela paciência e todos os ensinamentos, conversas e conselhos incorporados nesse projeto ou no dia-a-dia.

Aos professores Geraldo Zimbrão e Jorge Leão por agregar prestígio a esse trabalho, compondo a banca examinadora.

Aos demais professores do PESC e da COPPE, que trabalham diariamente para possibilitar um ambiente de idéias e colaboração, sem o qual esse trabalho não poderia ter sido concebido. Não poderia deixar de citar os professores Jano de Souza e Claudia Werner, pelas lições aprendidas e aplicadas dentro e fora da sala de aula, nem os professores Collier e Benevides que me apoiaram desde o início, me recomendando para o programa.

Aos alunos de doutorado da linha de banco de dados, uma tropa de elite composta por pessoas como Ricardo Barros, José Rodrigues Neto, Bruno Osiek, Wallace Anacleto, Adriana Vivacqua e Jonice Oliveira, que contribuíram de varias e importantes formas para que esse trabalho finalmente ficasse pronto.

Aos amigos Marcelo Alvim e a Isabella Almeida pela grande contribuição de idéias, conversas e bons conselhos.

Às minhas chefes na Petrobras: Ana Cláudia Pinheiro e Verônica Schara pelo apoio, incentivo e paciência; aos meus colegas petroleiros que, de variadas formas, ajudaram para que esse trabalho pudesse ser realizado: Milton Cherman, Fábio Heuseler, Bruno Rocha, Marcos Ferreira, Vinicius Gomes, Thiago Siqueira e Elio Duarte.

Aos meus amigos, incluindo todos os já citados, que torceram por mim e por esse trabalho.

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

## CLASSIFICAÇÃO DE TEXTO PARA MELHORIA NO ATENDIMENTO DE HELP DESK

João Alquéres de Avellar Menezes

Outubro/2009

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Este trabalho tem como objetivo a aplicação de técnicas de mineração de textos no processo de atendimento de chamados realizados pelos Help-desks das organizações. Tais técnicas são apresentadas e sua aplicabilidade para esse cenário é discutida juntamente com as melhores formas de aplicação. Um experimento foi projetado e realizado para medir a eficácia das idéias propostas, utilizando dados retirados do Help Desk de TI da Petrobras, uma das maiores empresas do Brasil.

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

## TEXT CLASSIFICATION FOR HELP DESK SUPPORT

João Alquéres de Avellar Menezes

October/2009

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

This work proposes the application of text mining techniques to improve the tracking and solving of issues and problems conducted at IT help desks. The industry best practices are studied and their implementations in a major corporation are reviewed. Data mining techniques used to achieve the proposed goal are presented along with the best ways to apply them in the business scenario. An experiment was designed to validate the proposal and to collect data regarding the potential gains of the proposal. The experiment was tested using data gathered from a real world IT help desk from Petrobras, one of the world leading industries in oil exploration.

## Sumário

Capítulo 1 – Introdução .....	1
1.1 Objetivos: .....	1
1.2 Organização: .....	3
Capítulo 2 – Caracterização do Problema.....	5
2.1 Gestão de Serviços de TI .....	7
2.2 O Sistema Lig-Ti.....	8
2.3 O chamado .....	9
2.4 Mesas de Atendimento.....	12
Capítulo 3 – Revisão Bibliográfica.....	16
3.1 Aprendizado de máquina .....	16
3.2 Mineração de dados e de textos .....	17
3.3 Classificação de textos.....	18
3.4 Algoritmos de classificação de textos.....	21
3.5 Otimização .....	25
Capítulo 4 – Metodologia de trabalho .....	29
4.1 Proposta de processo automatizado .....	31
4.2 Revisão da Solução .....	26
4.3 Classificação automática dos chamados .....	32
Capítulo 5 – Aplicação da proposta .....	37
5.1 A Arquitetura do classificador .....	37
5.2 Seleção de dados .....	39
5.3 Pré-Processamento dos textos.....	41
5.4 Validação dos resultados.....	42
5.5 O Experimento .....	43
5.5.1 Resultados experimentais da classificação C0:.....	44
5.5.2 Resultados experimentais da classificação C1 .....	45
5.5.3 Resultados experimentais da classificação C2.....	47
5.5.4 Resultados experimentais da classificação C3.....	49
5.6 A Implementação .....	50
5.6.1 Resultados obtidos .....	51
5.6.2 Ganhos Percebidos.....	54
Capítulo 6 – Conclusões .....	56
Referências Bibliográficas .....	58

## **Lista de Figuras:**

Figura 1 – O sistema Lig-Ti

Figura 2 – Diagrama de estados dos chamados

Figura 3 - Algoritmo kNN

Figura 4- Algoritmo naïve Bayes

Figura 5 - Algoritmo SVM

Figura 6 - Representação do processo atual de atendimento

Figura 7 – Exemplo de Atendimento de chamado

Figura 8 - Representação do novo processo de atendimento

Figura 9 – Aplicação de predição de tempo de atendimento.

Figura 10 - A arquitetura do classificador implementado



## **Lista de Tabelas:**

Tabela 1 - Tempo médio de designação e solução de chamados

Tabela 2 – Atributos de um chamado

Tabela 3 – Status de chamado

Tabela 4 – Distribuição de chamados no primeiro semestre de 2009

Tabela 5 – Palavras mais comuns

Tabela 6- Resultados do experimento com a etapa C0

Tabela 7- Resultados do experimento com a etapa C1

Tabela 8- Resultados do experimento com a etapa C2

Tabela 9- Resultados do experimento com a etapa C3

Tabela 10 - Resultados da etapa C0 da classificação seqüencial

Tabela 11 - Resultados da etapa C1 da classificação seqüencial

Tabela 12- Resultados da etapa C2 da classificação seqüencial

Tabela 13- Resultados da etapa C3 da classificação seqüencial

## **Lista de Símbolos/Nomenclatura**

ITIL – Information Technology Infrastructure Library

BW – Business Warehouse

TI – Tecnologia da Informação

kNN – k-nearest neighbors

SVM – Support Vector Machines

NB – Naive Bayes

# Capítulo 1 – Introdução

O antigo ditado popular que diz que tempo é dinheiro nunca foi tão verdadeiro como no mundo corporativo de hoje. Os gastos com operações de TI das grandes empresas são de milhões de dólares por dia. Com a finalidade de buscar reduções nos custos, os processos de otimização e automação de tarefas de manutenção para infraestrutura são muito procurados e estudados. (Weill et al, 2002)

A qualidade de dados em um negócio possui impactos em diversas áreas. Redman(1998) mostra como as organizações são afetadas, do topo ao chão de fábrica, pela qualidade dos dados que possuem sobre seus negócios. Dados incompletos ou errados podem levar, no nível operacional, a desperdício e retrabalho. No nível gerencial os impactos do uso de dados de baixa qualidade podem ser bem maiores, comprometendo os planos de médio e longo prazo da empresa, atrapalhando o posicionamento mercadológico, reduzindo a confiança dos clientes ou até alienando a alta administração.

Administrar dados confiáveis e permitir com que eles estejam disponíveis para consulta quando forem necessários é uma tarefa complexa, e são gastas grandes quantias de dinheiro anualmente para tornar a informação disponível e com qualidade para seus usuários. Sempre que um problema é encontrado nos sistemas ou bases de dados corporativas, ele precisa ser corrigido o mais rápido possível, de forma a mitigar os impactos causados por esse erro.

## 1.1 Objetivos:

Esse trabalho tem como objetivo estudar uma maneira de reduzir o tempo gasto no processo de correções de erros nos sistemas corporativos. Esse processo, que envolve

a identificação de um problema, a atribuição de responsabilidade pela correção e sua subsequente solução é apresentado e analisado, em seguida é feita uma proposta de melhoria, baseadas nos conceitos apresentados nesse trabalho.

No contexto dos negócios, os gerentes de TI comumente entendem que a melhor forma de melhorar os tempos de atendimento é atuando sobre a etapa do processo em que o solucionador interage, ou no tempo em que um chamado aguarda na fila para ser atendido. Medidas comuns incluem o aumento do número de recursos alocados para a atividade, ou remanejamento de prioridades. Esse trabalho propõe uma abordagem diferente: a redução do tempo de atendimento através da redução do tempo de atribuição de um problema para a pessoa ou a equipe responsável pela manutenção do sistema (ou módulo do sistema) afetado.

Este trabalho apresenta um estudo de um caso real, mostrando como o processo é realizado em uma grande empresa e descreve uma tentativa de economizar tempo de atendimento. Essa proposta consiste na substituição do processo atual, executado de forma manual, por uma versão automatizada baseada em técnicas de mineração de dados e classificação de textos. A situação atual do processo é estudada e são identificadas oportunidades de melhoria no processo.

O uso de um classificador de textos visa atribuir a um documento sua categoria. Se visualizarmos essas categorias como os grupos responsáveis pelos serviços de TI, encontrar o grupo responsável reduz-se, simplesmente a encontrar a classe de um texto. Esse trabalho apresenta uma comparação entre os tempos empregados nessa atividade no caso observado e o tempo que seria necessário para se encontrar a categoria de um texto. Tais tempos são estudados nesse trabalho, de forma a evidenciar os ganhos potenciais

resultantes da aplicação das idéias propostas.

Um experimento, usando dados reais, é apresentado para avaliar a proposta apresentada. Nesse experimento serão usados tanto para testes como para treinamentos dados de chamados registrados no sistema de atendimento em produção. Chamados encerrados e com sua classificação devidamente validada pelos especialistas no processo são usados como gabarito para o classificador.

Após o estudo desses resultados as idéias propostas são implementadas em uma aplicação para simular o processo automatizado. Os resultados obtidos são comparados com as informações coletadas sobre o processo manual, de forma obter medidas do desempenho do novo processo.

Este trabalho pretende provar que é possível, com uma alta taxa de sucesso, encontrar a mesa apropriada para um chamado e ainda registrar a redução no tempo total do atendimento graças a essa rápida classificação.

## **1.2 Organização:**

Esse trabalho está organizado em seis capítulos, apresentados a seguir:

Este primeiro capítulo consiste numa introdução ao trabalho, apresentação do problema e descrição sucinta dos objetivos.

O segundo capítulo caracteriza o problema abordado, descreve o processo de atendimento de chamados tal como ele foi definido e é aplicado atualmente.

O terceiro capítulo apresenta uma revisão dos conceitos empregados na tentativa de sanar o problema discutido.

No quarto capítulo é apresentada a proposta central do trabalho, a abordagem que

visa mitigar o problema, baseada na aplicação dos conceitos revisados no terceiro capítulo. É idealizada uma ferramenta que possibilitaria a substituição do processo atual pelo novo processo proposto.

O quinto capítulo discute o experimento projetado para aferir o potencial da solução, formas de avaliar a qualidade do experimento e uma análise sobre como os resultados podem ser comparados com o processo real. A partir desses resultados é descrito como foi construída a aplicação que habilita a aplicação do processo proposto nas estruturas de help desk das organizações.

Por fim, o sexto e último capítulo contém as conclusões extraídas do trabalho, discute os próximos passos e oportunidades de melhorias identificadas.

## Capítulo 2 – Caracterização do Problema

O estudo realizado nesse trabalho foi motivado pela observação dos problemas e oportunidades de melhoria do sistema de gestão que é atualmente utilizado pelo help desk da Petrobras. Esse capítulo descreve brevemente o funcionamento atual do processo de registro de problemas e seu subsequente tratamento.

A Petrobras, Petróleo Brasileiro S/A, é uma empresa brasileira de economia mista. Fundada no dia 3 de outubro de 1953 e sediada no Rio de Janeiro, opera hoje em 27 países, prioritariamente nas áreas de exploração, produção, refino, comercialização e transporte de petróleo e seus derivados. A empresa está em quarto lugar no ranking das maiores petrolíferas de capital aberto do mundo, é a terceira maior empresa do continente americano em valor de mercado e ocupa o sexto lugar entre as maiores empresas do mundo em valor de mercado.

Gerenciar os eventuais desvios e incidentes nos sistemas de Tecnologia de Informação de uma empresa desse porte é um trabalho que consome grande esforço, tempo e dinheiro. Números internos da Petrobras indicam que cerca de 500 analistas trabalham no atendimento dos mais de 77.000 usuários nos sistemas de TI, entre empregados e contratados. São registrados mais de 140.000 ligações e e-mails por mês. Se analisarmos somente os atendimentos relacionados à principal ferramenta de gestão integrada da empresa, o SAP R/3, tratamos de um conjunto de 40.000 ligações por mês.

Devido à quantidade massiva de dados registrados no sistema de acompanhamento, muitos deles classificados com um nível de segurança que impossibilitam sua publicação nesse trabalho, iremos tratar apenas de um subconjunto desses dados. Os critérios para seleção desse subconjunto são apresentados a seguir, bem

como considerações sobre como as conclusões encontradas podem ser generalizados para todo o universo de atendimento. O conjunto a ser analisado, escolhido principalmente por questões de familiaridade, acesso aos dados e aos especialistas no processo é a fração da estrutura de atendimento relacionada ao sistema de Data Warehouse corporativo da empresa, o BW.

O Business Warehouse (BW) é um sistema desenvolvido pela empresa alemã SAP - *Systeme, Anwendungen und Produkte in der Datenverarbeitung*. Esse sistema é responsável pela leitura de dados de diversos sistemas transacionais em uso da companhia para geração de diversos relatórios operacionais e gerenciais. Ele possui armazenados diversos terabytes de dados, recebendo diariamente novos dados de aplicações e assuntos variados, envolvendo os processos de compras, pagamentos, contabilidade, gestão de recursos humanos e muitos outros. Muitas outras informações sobre esse sistema e sua implantação na Petrobras podem ser encontradas em Pinheiro(2006).

Todos os chamados desse conjunto específico, em um intervalo de tempo de seis meses (entre Janeiro e Junho de 2009) foram analisados e os tempos totais de atendimento foram registrados na tabela a seguir. Esses tempos foram segmentados em duas atividades: Tempo total até o chamado ser atribuído ao grupo responsável pela sua conclusão e tempo total de atendimento por esse grupo:

Nível	Nº de chamados	Designação (hs)		T solução (hs)	
		Média	Desvio Pd.	Média	Desvio Pd.
N1	841	0:00	0	0:13	1,9
N2	379	1:42	8,5	2:59	10,0
N3	1574	3:47	12,0	28:17	42,3
<b>(média)</b>		<b>2:22</b>		<b>16:24</b>	

*Tabela 1 - Tempo médio de designação e solução de chamados*



Essa tabela indica que o tempo de designação dos chamados ao seu grupo responsável consome 15% do tempo total de atendimento.

## **2.1 Gestão de Serviços de TI**

O gerenciamento de serviços de TI tem por objetivo prover um serviço de TI alinhado às necessidades do negócio, buscando sempre uma redução de custos e tempos de atendimento a longo prazo. Um dos padrões mais conhecidos de gestão de serviços é descrito pela disciplina de Service Support do ITIL.

A Information Technology Infrastructure Library (ITIL) é uma biblioteca de boas práticas em desenvolvimento desde o final dos anos 80 pelo governo inglês. O ITIL busca promover a gestão com foco no cliente e na qualidade dos serviços de tecnologia da informação.

O Service Support é uma disciplina do ITIL cujo foco está na manutenção dos serviços de TI das companhias, de forma que eles possam ser usados para suportar os negócios da empresa. A interação dos usuários com o suporte se dá através de solicitações de mudanças, relatório de problemas e pedidos de esclarecimentos.

Service Desk é o instrumento básico para a gestão de serviços, conforme definido no ITIL. Sua função é centralizar todos os contatos entre usuários e os provedores de um serviço, atuando como ponto focal para registro de incidentes e para solicitações de serviços. Um Service desk é composto por diversos grupos de atendimento, ou mesas, com responsabilidades bem definidas e organizadas hierarquicamente de acordo com a complexidade das soluções envolvidas.

## 2.2 O Sistema Lig-Ti

O Sistema Lig-Ti é um sistema adquirido pela Petrobras e usado em toda a empresa para gerenciar os problemas, incidentes e solicitações de serviço referentes aos sistemas de TI corporativos. Seu funcionamento está descrito no Padrão de Procedimento de Gestão de Incidentes, um documento interno cujo conteúdo é descrito nessa sessão..

O Lig-Ti é o sistema aonde todas as solicitações e contatos feitos por usuários são registrados. Tais solicitações ou relatos de incidentes são representados no sistema através de um chamado.

Chamados são tratados pelas equipes responsáveis pelas ferramentas da empresa. Essas equipes são representadas no sistema pelos Grupos de atendimento.

Chamados e grupos de atendimento são os dois objetos mais importantes do sistema, e são descritos com mais detalhe a seguir.

LIG-TI Atendimento (Modify)

BR PETROBRAS LIG-TI Service Desk Modifica Chamado

Solicitante Controle

Chave CZL8 Cadastro País Estado (UF) RJ Lotacao TIC/ADS-TI/SCD

Nome JOAO ALQUERES DE AVELLAR Município RIO DE JANEIRO

Telefone 817 5579 Imovel CITIBANK Local 13º ANDAR

Chamado	Situacao	Categoria	Produto	Servico	Descricao
Click to Refresh					

Exibir Atualizar

Visualizacao TI-RIO Situacao Fechado Chamado HD0000012912214

Categoria\* SAP Prioridade Normal

Produto BW PETROBRAS Tipo de Chamado Incidente

Servico REATIVAR SENHA

Causa DESBLOQUEIO DE USUA

Descricao CHAVE BLOQUEADA NO BW. Grupo TI-RIO-HD-SAP PETROBRAS

Complemento Tecnico MARCELA CRISTINA DOS SANTOS

Transferir Chamado Exibir Transferência/Origem Lembretes Fechar Imprimir Chamado Salvar

LIG-Ti Atendimento (Modify)

---

**BR PETROBRAS**      LIG-Ti Service Desk      Modifica Chamado

---

Solicitante: Controle

Historico  
 17/2/2009 15:04:36 Y1WL  
 Realizado o desbloqueio da Chave no Sistema. Testes OK. Não foi necessário acesso remoto.

Anexos

Arquivo	Tamanho	Rótulo
anexo1		
anexo2		
anexo3		

Rastreamento      Nível de Serviço

Cod. Equipamento:      Origem: Telefone

Tipo de Demanda?: Cliente      Data Criacao: 17/2/2009 15:02:14

Prazo(min):      Designado as: 17/2/2009 15:02:14

Criado por: Y1WL      Resolvido as: 17/2/2009 15:04:35

Ultima Modificacao por: CADLOTE      Ultima Modificacao as: 19/2/2009 20:29:10

---

Visualizacao: TI-RIO      Situacao: Fechado      Chamado: HD0000012912214

Categoria\*: SAP      Prioridade: Normal

Produto: BW PETROBRAS      Tipo de Chamado: Incidente

Servico: REATIVAR SENHA

Causa: DESBLOQUEIO DE USUA

Descricao: CHAVE BLOQUEADA NO BW.      Grupo: TI-RIO-HD-SAP PETROBRAS

Complemento:      Tecnico: MARCELA CRISTINA DOS SANTOS

Transferir Chamado      Exibir Transferência/Origem      Lembretes      Fechar      Imprimir Chamado      Salvar

*Figura 1 – O sistema Lig-Ti*

## 2.3 O chamado

O chamado é o objeto que representa um problema relatado pelo usuário. Ele é criado quando é feito um contato com o service desk e um problema é registrado. Cada chamado possui um número, que é seu identificador primário e nunca é alterado durante todo o ciclo de existência do chamado.

O chamado possui os seguintes atributos:

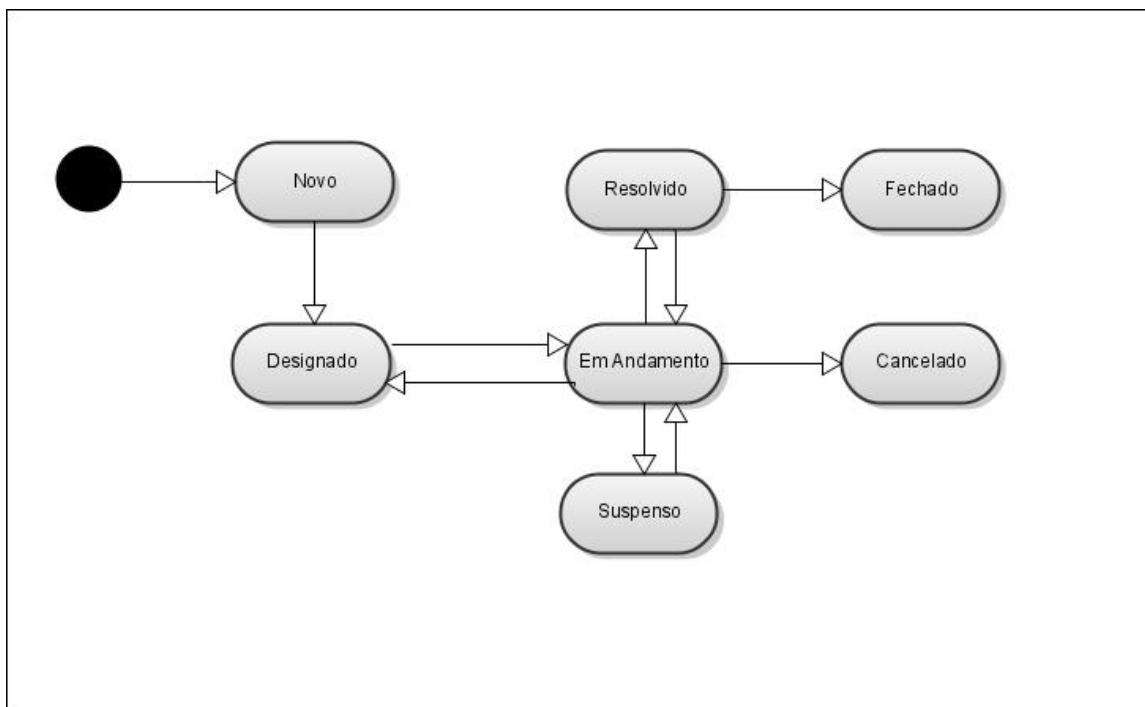
<b>Atributo</b>	<b>Tipo</b>	<b>Descrição</b>
Número	Número sequencial	O identificador único do chamado. É gerado automaticamente no momento que o chamado é criado e nunca é alterado.
Solicitante	Complexo	Nome e informações de contato do usuário que solicitou a abertura do chamado.
Grupo de atendimento.	Texto	Representa a equipe que responsável pelo chamado. Pode ser alterado durante o atendimento do chamado, simbolizando o repasse do problema para um grupo mais apropriado para o atendimento do mesmo.
Técnico	Texto	O técnico responsável pelo atendimento do chamado. Ao contrario do grupo, nem sempre está preenchido.
Data de Criação	Data	Data e hora de cadastro do chamado no sistema
Data de Solução	Data	Ultima data em que o chamado foi marcado como resolvido.
Data de Designação	Data	Ultima data em que o chamado recebe o estado de Designado.
Situação	Seleção em lista	O estado atual do chamado, conforme a tabela abaixo.
Produto	Seleção em lista	Indica o sistema ao qual o chamado se refere.
Descrição	Texto	Descrição do problema, em uma linha, cadastrada no momento de criação do chamado.
Complemento	Texto	Descrição mais detalhada do problema. Também é redigida no momento de aberta do chamado.
Histórico	Texto	O histórico é um conjunto de textos que descrevem todo o processo de atendimento de um chamado. Desde a análise do problema até os passos realizados para atendimento.

*Tabela 2 – Atributos de um chamado*

O chamado, ao ser criado quando o usuário entra em contato com a mesa N1, possui os seguintes estados, que são registrados no campo “Situação”.

<b>Estado</b>	<b>Descrição</b>
NOVO	Esse é o estado inicial dos chamados. Chamados recebem esse estado enquanto estão sendo criados. Após sua criação ele deve ser designado à mesa que irá tratar o chamado.
DESIGNADO	Sempre que um chamado é atribuído a uma mesa nova, ele recebe esse estado. Indica que o chamado está associado a uma mesa, mas não a um analista.
EM ANDAMENTO	Quando um analista recebe a notificação de que existe um chamado designado em sua mesa, ele acessa a ferramenta e pode assumir a responsabilidade de estudar o chamado, atribuindo esse estado a ele. O responsável pode designar um chamado para uma mesa de nível superior, ou devolvê-lo para o nível abaixo, caso julgue que o chamado não foi corretamente designado a sua mesa.
SUSPENSO	Caso seja necessário interromper temporariamente o atendimento de um chamado, ele pode receber esse estado, juntamente com uma descrição e previsão de solução da pendência.
RESOLVIDO	Quando um analista realiza o serviço ou a correção solicitada pelo usuário, ele pode encerrar o chamado marcando ele com esse estado.
CANCELADO	Em situações quando dois chamados tratam do mesmo problema, o chamado mais recente pode ser cancelado com a aprovação do usuário, de forma a evitar um fluxo excessivo de chamados para o grupo resolvidor. De forma semelhante a impossibilidade de contato com um usuário para esclarecimento de um problema, após diversas tentativas também pode levar ao cancelamento da solicitação.
FECHADO	Após um determinado período de tempo, os chamados resolvidos são fechados automaticamente. Isso indica que o problema foi de fato resolvido e que a solução está homologada. Este é um dos dois estados finais de um chamado.

*Tabela 3 – Status de chamado*



*Figura 2 – Diagrama de estados dos chamados*

De acordo com as normas de atendimento o campo de histórico do chamado deve relatar todos os detalhes importantes para se entender a causa de um problema relatado e a forma como ele foi tratado, mesmo nos casos em que o chamado é aberto apenas como cobertura para um problema já resolvido.

É considerada uma responsabilidade da equipe que fecha o registro, a revisão da categorização realizada, alterando-a caso necessário.

## **2.4 Mesas de Atendimento**

Mesas de atendimento são os componentes do Service Desk. Cada uma delas possui um determinado número de analistas responsáveis pelo tratamento dos chamados que são atribuídos às mesas. O termo grupo de atendimento também é usado para representar uma mesa.

As mesas estão organizadas em quatro níveis hierárquicos, de acordo com a sua função:

- 1º Nível de atendimento - Mesas de atendimento corporativo por telefone, web ou e-mail. As mesas do primeiro nível são as únicas mesas aonde novos chamados são criados. Os analistas dessas mesas são treinados para avaliar cada nova solicitação dos usuários, definir se a solução do problema faz parte do seu escopo de trabalho e, quando necessário, identificar qual é o grupo do próximo nível de atendimento mais apropriado para atender ao chamado. O tipo mais comum de atendimento realizado nesse nível é o de esclarecimento de dúvidas dos usuários.
- 2º Nível de atendimento - Mesas de serviços especializados ou corporativo sem desenvolvimento e sem infra-estrutura. As mesas desse nível de atendimento atendem as solicitações referentes aos grandes sistemas da empresa, como por exemplo, BW, Lotus Notes ou Oracle. Seus analistas não possuem conhecimentos dos processos de negócio, apenas nas questões técnicas das ferramentas. O tipo mais comum de atendimento realizados nesse nível é de execução de serviços de manutenção ou configuração do sistema.
- 3º Nível de atendimento - Mesas de serviços especializados. São os grupos responsáveis por ferramentas específicas e estão capacitados para tratar de problemas envolvendo as regras do negócio. As mesas de atendimento desse nível também interagem com as equipes de infra-estrutura solicitando alterações relativas à performance ou capacidade. Chamados

normalmente associados a esse nível incluem dúvidas sobre como os processos de negócio são mapeados no sistema.

- 4º Nível de atendimento - Mesas de serviços especializados com desenvolvimento. Essas mesas são compostas pelas equipes responsáveis pela manutenção e desenvolvimento de ferramentas usadas na empresa. Seu envolvimento é necessário quando o problema relatava envolver a correção de algum bug nos softwares, ou é encontrado algum problema na implementação dos requisitos de negócio. De todos os grupos que compõem o Help Desk esse é o único que não atua dedicadamente a essa função, seus profissionais são os responsáveis pelas ferramentas e recebem a atribuição adicional de suportar os usuários corporativos através do Help Desk.

Devido às características da equipe do N4, o processo prevê que os chamados nunca sejam efetivamente enviados a elas. Ao invés disso, quando um analista do N3 identifica a necessidade de uma correção pelo N4 ele solicita o trabalho através de uma tarefa para a mesa N4. Uma tarefa é um objeto semelhante a um chamado, com duas diferenças básicas: Uma tarefa sempre referencia um chamado atendido pelo suporte N3 e ela só existe entre sua mesa criadora e uma mesa N4 correspondente.

Além dessas atribuições específicas, todos os grupos possuem algumas responsabilidades comuns. Todos os grupos devem avaliar a pertinência do chamado, caso identificarem que o grupo foi incorretamente atribuído, devem devolvê-lo ao nível anterior que os encaminhou registrando os motivos. No caso



de identificarem que a solução está além da sua competência, devem encaminhar para o nível seguinte, registrando a análise que já tenha sido feita para apoiar o atendimento e evitar retrabalho.

Todos os analistas possuem acesso a uma base de conhecimentos que pode ser acessada para suporte no caso de dúvidas sobre a responsabilidade e o escopo de uma determinada solução.

## Capítulo 3 – Revisão Bibliográfica

Nesse capítulo são apresentados os conceitos que foram estudados e aplicados nessa tentativa de solucionar o problema descrito no capítulo 2.

### 3.1 Aprendizado de maquina

Aprendizado de maquina, ou Machine Learning, é a disciplina que estuda formas de habilitar computadores a tomarem decisões diante de situações novas, sem que estejam especificamente programados para reagir a um determinado estímulo. Aplicações comuns visam o reconhecimento de padrões e a tomada de decisões a partir de dados históricos.

Sua utilização tem tido um interesse crescente nos últimos anos, principalmente devido à crescente disponibilidade de documentos nos formato digital e a subsequente necessidade de organizá-los. (Sebastiani, 2002).

Tipicamente os algoritmos de aprendizado de maquina trabalham aplicando um processo que “estuda” um conjunto de dados e produz um modelo que armazena o conhecimento adquirido. Essa abordagem representou uma evolução importante em relação à forma como sistemas computacionais poderiam identificar ou tomar uma decisão: Ao invés de regras bem definidas, conjuntos de treinamento mais dinâmicos podem ser usados. Esses conjuntos podem facilmente ser estendidos e incorporar novas informações. Outra vantagem importante é que num sistema de regras existe a necessidade de especialistas definirem claramente cada classe. Tal abordagem requer muito trabalho para ampliar um conjunto de classes, uma vez que todas as novas classes precisam ter suas regras definidas e deve-se revisar as regras antigas, para evitar conflitos. Uma abordagem baseada em conjuntos de treinamento pode ser estendida sem

que haja a necessidade de se definir um novo conjunto de regras.

### **3.2 Mineração de dados e de textos**

Mineração de dados (Data mining) é o nome dado a atividade que consiste na procura de informações em um grande conjunto de dados (Han e Kamber, 2001). Essa atividade foi assim batizada em analogia a atividades dos mineradores de ouro, que trabalhavam procurando e separando uma pequena quantidade de mineral precioso dos grandes volumes de minérios sem valor.

A Mineração de textos (Text mining) pode ser definida como a atividade de recuperar e extrair informações úteis em grandes volumes textuais, escritos em linguagem natural, a fim de descobrir conhecimento inovador nos textos.

Historicamente, a importância dessa área ganhou impulso a partir da década de 90, com o crescimento do armazenamento digital, da internet e dos mecanismos de busca. Ao mesmo tempo, analistas começaram a perceber a ausência de ferramentas de mineração de dados para lidar com o ambiente de informações não estruturadas. (Lopes 2004).

Como a forma mais natural de armazenar informações é o texto, acredita-se que Text mining tenha um grande potencial no cenário corporativo. Em Furtado(004), são listados estudos que apontam que até 80% de informações em uma empresa estão em bases textuais. A natureza não estruturada dos textos faz com que a mineração de textos seja considerada uma tarefa um pouco mais complexa que mineração de dados estruturados. Algumas aplicações típicas de Mineração de texto são apresentadas a seguir:

- Predição: A descoberta de atributos desconhecidos.
- Regressão: A descoberta de tendências e dependências entre os dados.
- Agrupamento: A obtenção de conjuntos com características similares, como o assunto dos textos.
- Categorização: A descoberta de um grupo ao qual o elemento pertence.
- Análise de associações: Identificação de relacionamentos entre textos distintos.

### **3.3 Classificação de textos**

Classificação (ou categorização) de textos é a área de conhecimento que estuda formas de classificar um conjunto de textos em um conjunto de categorias previamente conhecidas.

RIZZI et al. (2000) afirmam que a categorização de textos é uma técnica utilizada para classificar um conjunto de documentos em uma ou mais categorias existentes. Ela é geralmente utilizada para classificar mensagens, notícias, resumos e publicações. A categorização também pode ser utilizada para organizar e filtrar informações. Essa capacidade faz com que esta técnica possa ser aplicada em empresas, contribuindo no processo de coleta, análise e distribuição de informações e, conseqüentemente, na gestão e na estratégia competitiva de uma empresa.

A categorização automática de um texto é possibilitada através de um modelo que indica quais são as classes a que o texto pertence a partir dos elementos que o compõe. Tal modelo é chamado de classificador.

A seguinte definição formal foi adotada para descrever um classificador: Um

classificador é uma função  $F(x_1, x_2, x_3, \dots, x_n)$  que associa a um objeto, representado por um conjunto de atributos  $x_1, x_2, \dots, x_n$  uma classe  $c_n$  pertencente a um domínio.  $C = \{c_1, c_2, \dots, c_n\}$ . (Dumais et AL, 1998).

Garantir a precisão do método de categorização, em qualquer que seja o contexto de classificação, e a redução do esforço humano para gerar bases pré-classificadas para treinar os categorizadores são ainda alvo de muitas pesquisas (SEBASTIANI, 2006) e existem diversos métodos e algoritmos para categorização de textos que trabalham de formas bastante diferentes em uso atualmente.

Existem métodos que categorizam textos em apenas uma categoria (single-label), enquanto que outros permitem que textos possam ser classificados em mais de uma categoria (multi-label). Nesses métodos é comum que seja gerado também um grau de pertinência do documento a cada uma das categorias para as quais ele foi classificado (LEWIS, 1998). Conforme discutido nesse capítulo, devido à natureza do caso estudado nesse trabalho, o foco será dado aos algoritmos de classificação simples.

Categorizadores podem ser baseados em diferentes abordagens, as mais comuns são:

- Categorizadores baseados em modelos probabilísticos.

São os categorizadores que utilizam modelos probabilísticos e baseiam-se no princípio do ordenamento de probabilidade (ZHANG, BERG et al., 2006). Estes métodos seguem a abordagem de ordenar os documentos baseados na probabilidade da relevância com relação à consulta baseados na distribuição estatística dos termos nos textos. Como exemplos, podemos citar a abordagem de Naive-Bayes e os Modelos de Markov Escondidos (HMM).

- Categorizadores baseados em árvores de decisão.

As árvores de decisão consistem de nós que representam os atributos, arcos que recebem os valores possíveis para estes atributos e folhas, que representam as diferentes classes do domínio. São muito utilizadas na construção de categorizadores por serem consideradas formas simples de representar o conhecimento.

- Categorizadores de força bruta.

Esses classificadores recebem esse nome, porque diferentemente de outras categorias, utilizam-se os dados de treinamento para verificar quais são os documentos mais similares ao novo documento (ZHANG, BERG et al., 2006). Não é criado um modelo. Nos outros tipos de categorizadores encontrar o modelo é lento, mas depois de encontrado, a identificação da categoria a qual um novo documento é mais rápido (categorizadores apressados). Exemplos deste tipo de categorizador são o Case-based Reasoning - Raciocínio Baseado em Casos (CBR) e o k-vizinhos próximos (KNN).

- Categorizadores Baseados em Inteligência Computacional

São categorizadores de aprendizado supervisionado que utilizam técnicas como Redes Neurais e Algoritmos Genéticos para obterem seus resultados. Tais técnicas são inspiradas pelos mecanismos e estruturas do cérebro humano. Já na utilização de algoritmos genéticos como utilizado em LOUREIRO, MARGOTO et al.(2005) são automatizadas as buscas por valores de parâmetros que maximizam o desempenho do categorizador gerado. Podemos citar Partical Swarm Optimization (PSO) que é baseado no comportamento social do cardume

de peixes ou de aves, como um dos muitos que vêm despontando na utilização em categorização de textos (YU, WANG et al., 2007).

### **3.4 Algoritmos de classificação de textos**

Uma vez apresentadas as abordagens mais comuns para classificação, serão listados alguns algoritmos específicos, que implementam essas abordagens. Os algoritmos apresentados aqui serão usados e comparados no quinto capítulo que trata da implementação

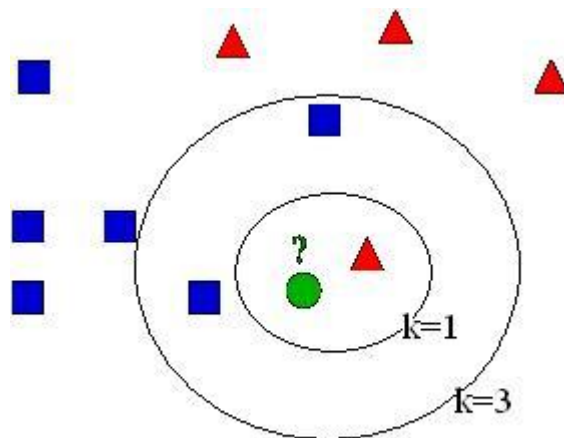
#### **3.4.1 Algoritmo dos K vizinhos mais próximos (kNN)**

O algoritmo dos k-vizinhos mais próximos, proposto em 1967, é um algoritmo de classificação baseada em similaridade (COVER e HART, 2008). É referenciado, neste trabalho, como o kNN, e tem sido largamente usado para resolver problemas de reconhecimento de padrão e classificação de documentos.

A base do algoritmo de kNN é uma função que mede a similaridade entre dois objetos, usando seus atributos. Partindo dessa medida, é possível comparar novos documentos com uma base de treinamento e encontrar documentos classificados que sejam semelhantes ao novo. Existem duas regras de classificação básicas: maioria na votação e peso pela distância. Na primeira, cada elemento tem uma influência igual, a classe escolhida é aquela que possuir mais representantes entre o kNN. Na abordagem por peso pela distância cada k-vizinho tem um peso inversamente proporcional à sua distância

Quando o k escolhido possui o valor unitário, o processo de classificação consiste simplesmente na busca pelo documento mais semelhante e replicar sua classificação.

Num caso aonde o valor de  $k$  seja igual ao tamanho do conjunto de treinamento, a classe predominante nesse conjunto determinará a classe do novo documento. A seleção do valor de  $k$  ideal é o objetivo de vários estudos, entre eles Friedman(1994) e Hastie e Tibshirani(1996), esses estudos apontam que a forma mais simples é através de experimentação e validação cruzada, quando possível. Outra preocupação comum no uso do kNN é a seleção de atributos. Para melhores resultados é desejável que se use poucos atributos muito significativos. Na classificação de textos, tal seleção pode ser vista como a seleção de palavras mais importantes em relação a palavras comuns, que podem ocorrer em qualquer classe, gerando ruído na classificação.



*Figura 3 - Algoritmo kNN*

A figura acima ilustra o funcionamento do algoritmo kNN. Ao se tentar classificar o objeto representado pelo círculo, o algoritmo irá prever a classe triângulo, quando parâmetro  $k$  for igual a 1, ou a classe quadrado quando  $K$  for igual a 3.

### **3.4.2 Naive Bayes**

Redes bayesianas são grafos acíclicos dirigidos que representam dependências entre variáveis em um modelo probabilístico. Esta abordagem representa uma boa

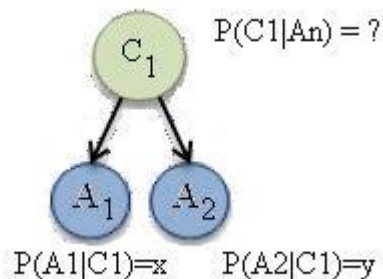


estratégia para lidar com problemas que tratam incertezas, onde conclusões não podem ser construídas apenas do conhecimento prévio a respeito do problema.

O algoritmo naive (ingênuo) Bayes é baseado na construção de um modelo probabilístico modelado como uma rede bayesiana. Essa modelagem se consiste no mapeamento das classes e dos atributos que as compõem. Aos relacionamentos são atribuídos probabilidades de que o atributo indique uma determinada classe.

O algoritmo foi batizado dessa forma por aliar o teorema de Bayes com a ingênua premissa de que os atributos são condicionalmente independentes, isto é, atributos de uma determinada classe não são informativos sobre os demais.

Apesar desta premissa ingênua e simplista, os classificadores de Naives Bayes possuem em desempenho satisfatório em muitas tarefas de classificação. (Chakrabarti, 2002)



*Figura 4- Algoritmo naive Bayes*

A figura acima ilustra um modelo ilustrativo do funcionamento de uma rede bayesiana. É conhecida a probabilidade de dois atributos, A1 e A2 fazerem parte de uma classe chamada de C1. O que deve ser determinado para a solução é a probabilidade do elemento fazer da parte da classe C1 dado que ele possui como atributos os valores A1 e A2.

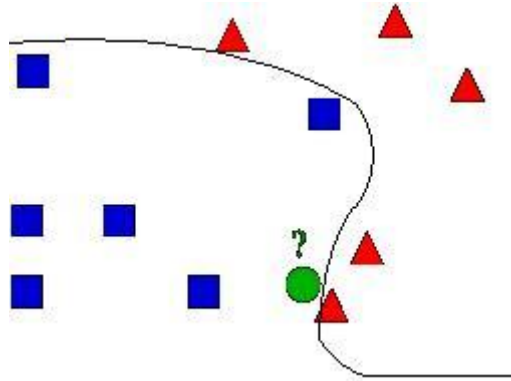
### 3.4.3 Máquina de Suporte a Vetor (SVM)

Support Vector Machines (SVM) consiste numa técnica de aprendizado para problemas de reconhecimento de padrão. A técnica vem sendo usada com grande sucesso em problemas de categorização de texto (Vapnik, 1992). O método SVM é essencialmente uma abordagem geométrica para o problema de classificação: Cada documento de um conjunto de treinamento pode ser interpretado como um ponto em um espaço multidimensional. Pontos pertencentes a uma classe são agrupados de acordo com funções que descrevem esses conjuntos. O treinamento do SVM consiste na busca de hiperplanos que dividem tais conjuntos.

Nessa abordagem geométrica, podem existir infinitos hiperplanos para separar dois conjuntos de pontos linearmente separáveis no espaço. Logicamente, o melhor plano para descrever a divisão das classes é aquele que reduz ao máximo a possibilidade de erro no caso de uma nova classificação. Esses hiperplanos podem ser encontrados usando os pontos mais próximos, que são conhecidos como vetores de suporte (support vectors). Esses pontos, ou vetores são encontrados com a aplicação de uma técnica chamada de Maximização Estrutural de Risco, que visa minimizar a generalização do erro ao invés de utilizar erros empíricos no treinamento dos dados. (SILVA e FAGUNDES, 2007).

As diferentes implementações do algoritmo SVM podem recorrer a diversos tipos de funções diferentes na procura de formar e descrever os hiperplanos a partir dos conjuntos de treinamento. A função escolhida para uma determinada implementação é chamada de kernel do algoritmo. Yang e Liu (1999) apontam, após um extenso estudo

comparativo que para tratar de problemas de classificação de texto os melhores resultados são obtidos com a escolha de funções lineares devidamente parametrizadas.



*Figura 5 - Algoritmo SVM*

A figura acima ilustra uma função que divide um plano bi dimensional em duas partições que correspondem às classes quadrado e triângulo. O ponto a ser classificado, representado com um círculo, será considerado da classe quadrado após a avaliação do classificador.

### **3.5 Otimização**

Otimização é um campo de pesquisa que estuda de formas a se encontra os melhores elementos para o atendimento de algum critério, dentro de um conjunto de alternativas. Matematicamente falando, é a escolha de um conjunto de parâmetros que maximiza (ou minimiza, de acordo com o que se deseja encontrar) o valor de uma função.

Um dos algoritmos mais simples de otimização é o algoritmo de Hill Climbing. Hill Climbing é um método que visa encontrar um caminho para a maximização de uma determinada função. Ele possui esse nome em analogia a trabalho de um alpinista, que procura um caminho mais curto, ou que requer menor esforço, para o cume de uma

montanha.

O Hill Climbing trabalha de forma iterativa, gerando constantemente um palpite sobre qual é o próximo passo a ser adotado, analisando as opções possíveis.

Russel e Norvig (1995) relatam que esse método é bastante utilizado, devido a sua simplicidade e baixo custo computacional. Apesar disso, esse método, em sua forma básica, não garante que seu resultado final irá apresentar o melhor valor possível para a função avaliada. Caso essa função possua diferentes pontos de máximos locais, a aplicação do algoritmo pode levar a um desses máximos locais e não ao máximo global.

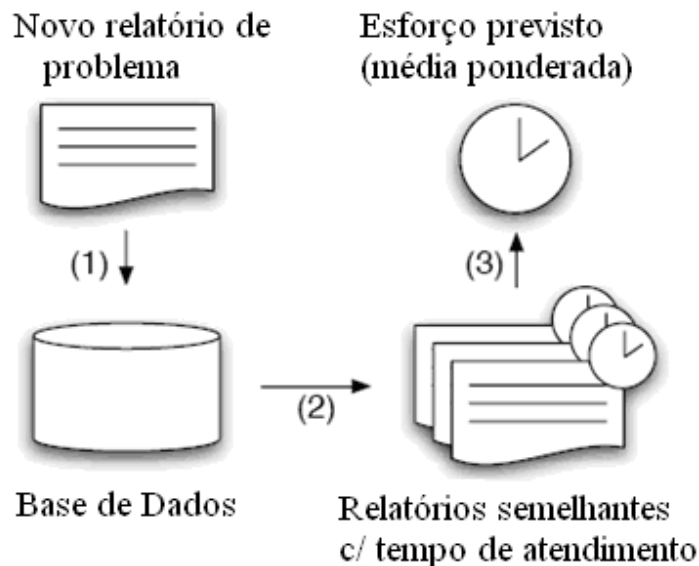
A aplicação de técnicas de otimização na busca dos melhores resultados de tarefas de classificação foi o objeto de estudos de Eads et AL (2002). Uma abordagem semelhante pode ser vista em FRIEDRICHS e IGEL (2005), que buscaram os melhores parâmetros para aplicar a um algoritmo de classificação SVM, utilizando técnicas de otimização para testar um grande número de combinações.

### **3.6 Trabalhos Semelhantes e Revisão da Solução**

Cubanic e Murphy (2004) propuseram uma forma de melhorar o tempo de atribuição dos relatos de bugs para os responsáveis pelos respectivos projetos. Essa abordagem é baseada na classificação dos textos que compõe os relatos visando atribuir automaticamente os relatos aos seus responsáveis. Devido à natureza dos projetos analisados, com diversos colaboradores com responsabilidades freqüentemente sobrepostas, e com grande número de relatos enviados pelos usuários com descrições incompletas ou erradas, a abordagem proposta não foi muito bem sucedida. O esquema de classificação aplicado resultou numa eficácia de apenas 30%. Apesar dos resultados fracos, os autores notaram que qualquer redução na carga de trabalho dos especialistas

encarregados de fazer a triagem dos relatos produzia um impacto positivo no projeto e notaram o grande potencial da aplicação.

Ao estudar formas de melhorar as estimativas sobre o tempo que correções de software requerem, Weiss (2007) propôs a comparação entre um novo esforço de desenvolvimento e outros, já realizados. Tal comparação, feita usando as informações textuais do relatório do problema, buscava uma lista de problemas semelhantes, cujos tempos de solução já eram conhecidos. A partir desses problemas já resolvidos e de seus respectivos tempos de solução é gerada a previsão de quanto tempo a solução do novo problema vai demandar. Essas idéias foram testadas em um experimento usando chamados submetidos ao repositório de problemas e solicitações de melhorias do projeto JBOSS. Segundo Weiss, tal avaliação permitiu a obtenção de estimativas melhores que as usadas anteriormente pela equipe do projeto. As novas estimativas tiveram um impacto positivo no trabalho, ajudando nos processos de priorização das correções.



*Figura 6 – Aplicação de previsão de tempo de atendimento.*

Anvik 2006 propôs uma forma de melhorar a qualidade de correções de problemas baseada na idéia de que o membro da equipe mais indicado para solucionar um problema é aquele que possui mais experiência com problemas semelhantes. Nesse estudo técnicas de categorização de textos são apresentadas como uma ferramenta importante para identificar semelhanças entre os relatórios de problemas. A abordagem consiste no levantamento de problemas semelhantes e subsequente levantamento de quem é o membro da equipe mais envolvido na solução desses problemas

Nesse estudo, foram analisados dados dos repositórios de bugs de dois projetos grandes: Mozilla Firefox e Eclipse. Foi identificado que a etapa de triagem dos novos relatórios de problemas poderia ter um ganho substancial de tempo, se eles pudessem ser diretamente encaminhados para o analista com mais experiência em resolver problemas semelhantes ao novo. Nesse trabalho, Anvik analisou diferentes algoritmos para classificação de textos e comparou os resultados para cada conjunto de dados estudados.

## Capítulo 4 – Metodologia de trabalho

Com base nos conceitos revisados no capítulo anterior, uma proposta de solução para o problema foi formulada. Esse capítulo fornece uma descrição detalhada do trabalho idealizado, assim como um histórico sobre o caminho percorrido.

Conforme visto no segundo capítulo, uma parte expressiva do tempo de atendimento dos chamados é gasto na alocação do chamado ao grupo de atendimento mais apropriado para seu atendimento. Uma das formas de se reduzir o tempo total de atendimento é reduzindo o tempo gasto nessa atividade específica. A proposta aqui apresentada consiste na substituição do processo manual e hierárquico de encaminhamento dos chamados por um processo automatizado, de forma a realizar de forma instantânea a designação de um chamado para a mesa de atendimento.

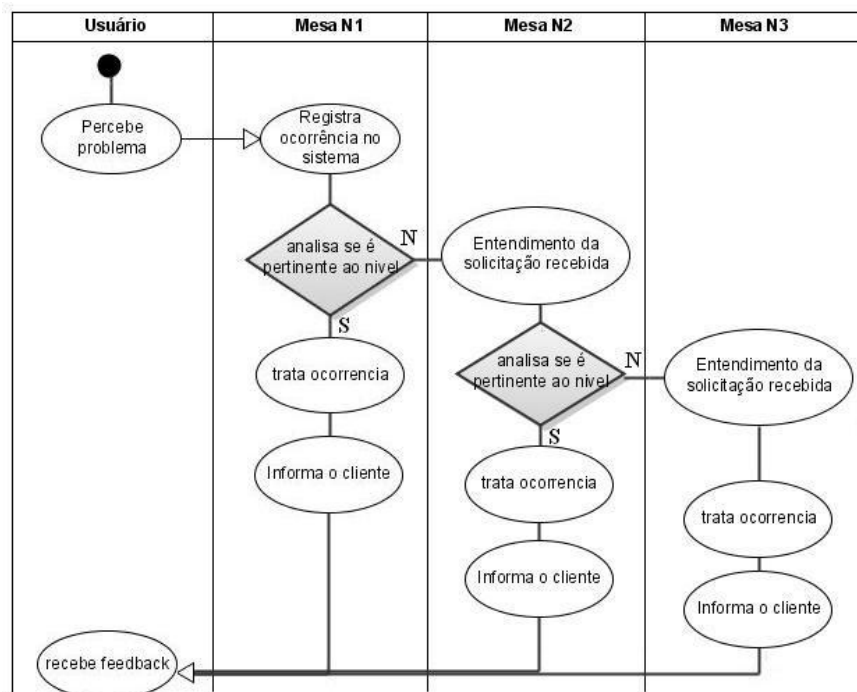
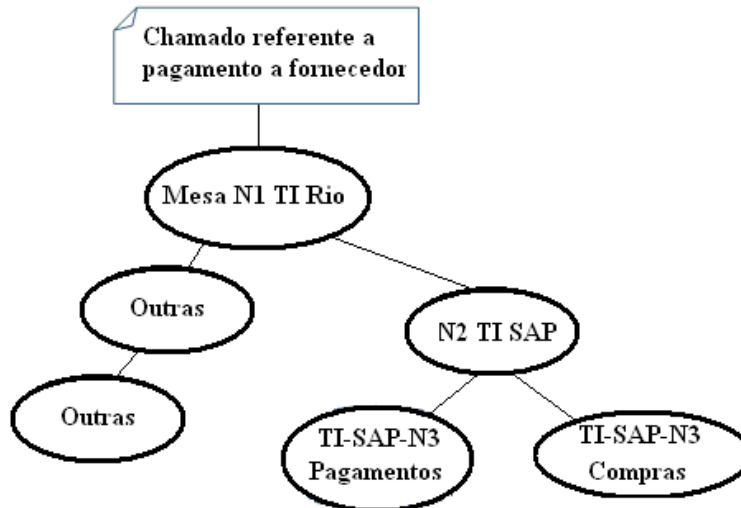


Figura 7- Representação do processo atual de atendimento

O processo em uso atualmente para busca da mesa apropriada para atendimento de um chamado foi desenhado com a intenção de possibilitar que vários analistas, cada um com conhecimentos limitados sobre as ferramentas e soluções de TI em uso, possam atuar em conjunto para encaminhar um problema para o grupo responsável. Um analista não precisa conhecer exatamente a mesa mais indicada para o tratamento de um problema, basta apenas saber identificar qual é o assunto ao qual o chamado se refere e encaminhar para a mesa do nível seguinte que trata desse assunto. A imagem abaixo mostra um exemplo de encaminhamento de chamado através do processo manual.



*Figura 8 – Exemplo de Atendimento de chamado*

A figura ilustra um caso em que é registrado, via telefone, um problema referente a um fornecedor. Esse fornecedor não recebeu seu pagamento no prazo estipulado e entrou em contato como responsável pela compra solicitado informações. O responsável, desconhecendo por que o pagamento não foi realizado, abre um chamado solicitando esclarecimentos. O chamado é criado pelo analista de uma mesa N1, que não tem conhecimentos suficientes sobre o processo para indicar se o problema deve ser visto pela

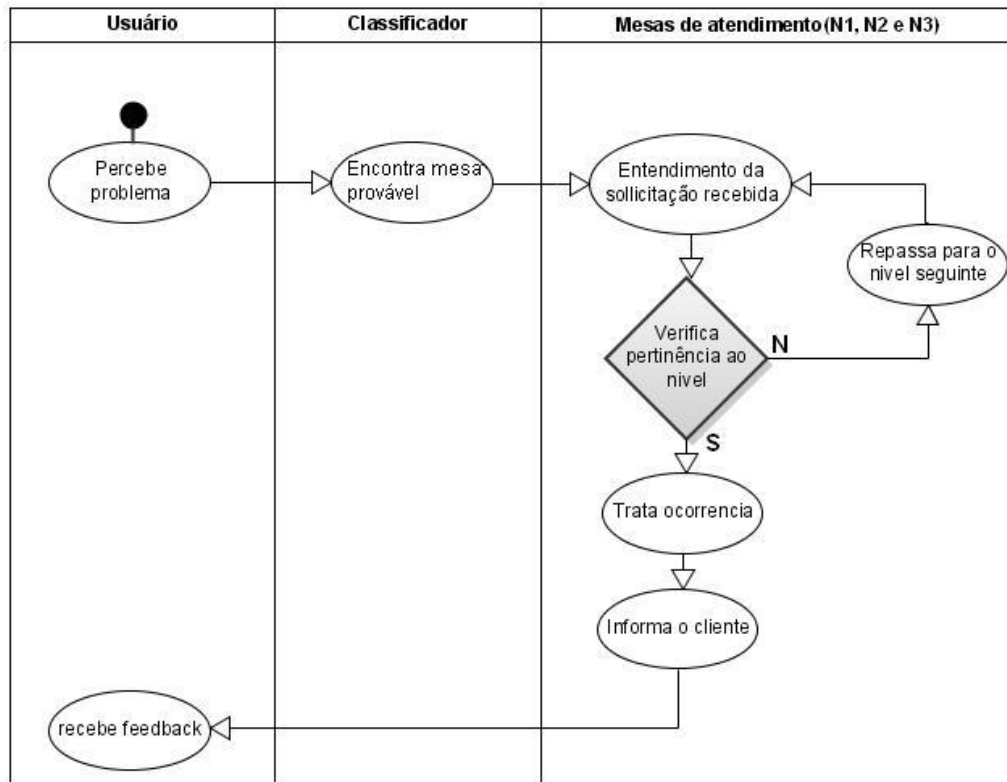


equipe que trata do processo de compras ou do processo de pagamento, ele simplesmente sabe que esse tipo de problema é relativo ao sistema integrado SAP. Dessa forma ele encaminha para verificação para esse grupo N2. O analista N2 irá, provavelmente, retornar o contato com o usuário e solicitar mais informações sobre a compra, para que ele possa entender melhor o problema e encaminhar para o suporte N3 apropriado.

#### **4.1 Proposta de processo automatizado**

Após a observação e estudo do processo atualmente utilizado, é apresentada uma nova versão desse processo que consiste na identificação automática de qual é o grupo final de um chamado utilizando apenas as informações disponíveis no momento de sua criação, isto é, o texto informado pelo usuário no momento em que ele entra em contato com o help desk. Se entendermos que as mesas de atendimento são as categorias as quais um chamado pode pertencer, podemos analisar as informações textuais que compõem o chamado e aplicar as técnicas de classificação para obter qual é a mesa que deverá atender esse chamado.

Essa classificação habilita o encaminhamento do chamado diretamente para o grupo responsável e evita o seu tramite pelos níveis intermediários do sistema. Essa abordagem oferece duas vantagens claras: Em primeiro lugar, o tempo de atendimento de cada chamado será impactado positivamente com a eliminação do tempo gasto na atividade de atribuição dos chamados aos seus grupos finais. Adicionalmente, haverá um impacto positivo na carga de trabalho das mesas de atendimento de nível mais baixo, pois elas deixarão de atuar como roteadores de chamados e passarão a se concentrar apenas nas atividades de atendimento dos chamados pertinentes a elas.



*Figura 9- Representação do novo processo de atendimento*

Retornando ao exemplo descrito anteriormente, ao analisar as palavras relatadas pelo usuário ao informar o problema, um classificador textual poderia associar esse problema diretamente à mesa N3 que recebe tipicamente esse tipo de chamados.

### 4.3 Classificação automática dos chamados

O processo de classificação automática de chamados consiste na aplicação de técnicas de classificação de texto para direcionar chamados automaticamente para o grupo de atendimento apropriado. A atividade de classificação textual pode ser dividida em três etapas (Neves, 2001):

1. Preparação dos dados - Essa etapa consiste no levantamento dos dados a serem usados na composição do conjunto de treinamento do classificador.

Adicionalmente, esses dados podem sofrer uma filtragem ou tratamento para eliminação de caracteres ou palavras irrelevantes, como preposições ou artigos, para reduzir a carga sobre o trabalho de classificação. Essas palavras irrelevantes são conhecidas como stopwords.

2. Representação - Essa etapa consiste na representação dos textos em um formato adequado para ser interpretado pelos algoritmos de classificação.
3. Classificação - Consiste na análise de novos chamados a partir da base de treinamento para atribuição das classes.

Além dessas três etapas, será realizada após a classificação, uma avaliação dos resultados obtidos, que consistirá na coleta dos indicadores sobre o desempenho do classificador.

#### **4.3.1 Os algoritmos de classificação utilizados**

Em seu trabalho de predição de tempos de atendimento de relatos de “bugs” Weiss et al (2007) utilizaram o algoritmo de kNN. Entre os motivos apresentados, é destacada a simplicidade e a flexibilidade do algoritmo como fatores decisivos para a sua escolha. Cubranic e Murphy (2004), adotaram em seu trabalho de triagem de “bugs” uma abordagem probabilística, utilizando o algoritmo de naive Bayes. Conforme apresentado nesse trabalho, o algoritmo foi escolhido devido aos bons resultados obtidos em testes comparativos e pela boa performance em classificações entre diversas classes. Anvik et al (2006) por sua vez, optaram pela escolha do algoritmo SVM em seu trabalho, devido ao ótimo desempenho desse algoritmo para problemas de classificação textual, conforme relatado em Joachins (1998).

Camargo (2007) revisou e testou diversos algoritmos de classificação de texto utilizados conjuntos textuais em português, anotando e comparando o desempenho deles. Neste trabalho o algoritmo de naive Bayes é apontado como apresentado o melhor resultado e é ressaltada a importância do pré processamento dos textos, excluindo stopwords para a obtenção de melhores resultados.

Para a execução das tarefas de classificação de textos empregadas no experimento contido nesse trabalho foram utilizadas todas as técnicas mencionadas nessa seção: kNN, Naive Bayes e SVM. Essa decisão visa coletar os resultados desses diferentes algoritmos com a intenção de identificar qual é o mais apropriado para esse domínio específico.

#### **4.3.2 Avaliação dos classificadores**

Após a execução do processo de classificação com os dados do conjunto de testes são produzidos indicadores sobre o processo de classificação. Tais indicadores têm como função aferir, através de diferentes aspectos, a qualidade da classificação.

A medida principal de avaliação do desempenho de classificadores é a acurácia, definida como o número total de documentos corretamente classificados dividido pelo número total de exemplos classificados. Outras medidas incluem a precisão (precision), a cobertura (recall), também chamada de revocação e a medida F (F-measure).

A **precisão** é a medida que expressa a fidedignidade de um resultado. No contexto desse experimento, a precisão indica quantos chamados classificados para uma determinada classe são realmente dessa classe. **Cobertura** é a medida que expressa a completude de um resultado. Nesse contexto, cobertura indica dentre todos os chamados que fazem parte de uma classe, quantos foram efetivamente encontrados pelo classificador. Devido à natureza complementar dessas duas medidas distintas foi

introduzida a **medida F**, que expressa com apenas um valor a média harmônica da precisão e da cobertura.

Para medir os impactos de um acerto e de um erro na atividade de classificação em relação a todo o processo de classificação automática de chamados, uma nova métrica será adotada, o custo de repasse. Essa medida pretende expressar o impacto que uma classificação errada pode causar no processo. Ela representa o prejuízo de tempo que um chamado, classificado em uma mesa incorreta, causará até que ele seja atribuído a mesa em que pertence realmente.

O **custo de repasse** será definido como a distância, na hierarquia, entre a mesa que ao qual o chamado foi atribuído e a mesa correta. Dessa forma, quando um chamado for corretamente atribuído o custo do erro será igual a zero. Um chamado incorretamente atribuído ao nível anterior, ou seguinte, terá o custo total de um. Quando um chamado for atribuído a uma mesa de terceiro nível incorreta, o custo de encaminhá-lo para a mesa correta é igual a dois. Isso ocorre por que os chamados são sempre atribuídos a níveis superiores ou inferiores. Um analista deveria devolver um chamado incorretamente para o nível superior, que por sua vez se encarregaria de atribuí-lo para a mesa correta.

Essa forma de calcular o custo permite a comparação entre o processo atual e o proposto. No processo atual, todos os novos chamados são automaticamente atribuídos a alguma mesa no nível 1. Após análise os chamados pertinentes aos níveis superiores serão atribuídos pelos analistas. Isso permite que seja atribuído a um conjunto de chamados um custo conforme as regras propostas acima. No processo atual, todos os chamados pertinentes ao N1 possuem custo 0, pois são criados na mesa aonde serão resolvidos. Analogamente, todos os chamados N2 possuirão custo 1, os chamados N3

possuirão custo 2. Como os chamados solucionados pelas as equipes N4 nunca são diretamente atribuídos a essas mesas, o custo para transmiti-los às mesas apropriadas também é 2, assim como todos os chamados N3. Num caso ideal de classificação automática, aonde todos os chamados são corretamente atribuídos aos seus respectivos destinos, o custo calculado seria 0. Esses dois valores estipulam um limite superior e inferior um processo de classificação automático.

## Capítulo 5 – Aplicação da proposta

Com a intenção de aplicar a o uso de técnicas de classificação de texto para descoberta do grupo de atendimento de um chamado, foi implementado um classificador usando os conceitos descritos na revisão bibliográfica e o processo apresentado no capítulo anterior. Um simples experimento foi desenhado para testar a aplicabilidade da solução proposta e por fim um classificador de chamados foi construído para verificar o processo apresentado, bem como as idéias usadas em seu projeto e validar os resultados obtidos. Também são apresentados objetivos e métricas para medir a eficiência da solução desenvolvida.

Para realização do experimento foi utilizado o Weka, um conjunto de ferramentas para aprendizado de maquina projetado para auxiliar na aplicação de técnicas de aprendizado para diversos problemas. WEKA é um acrônimo para Waikato Environment for Knowledge Analysis (Ambiente para análise de conhecimento de Waikato). Desenvolvido na Universidade de Waikato, na Nova Zelândia. Ele foi desenhado com a intenção de ser fácil de usar e permitir que um público maior de pesquisadores e cientistas tenha acesso a ferramentas de aprendizado de maquina. Sua aplicação possibilita a resolução de variados problemas relacionados a diferentes áreas de Data Mining, como classificação, clusterização, seleção de atributos e geração de gráficos. Garner(1995) fornece mais informações sobre o Weka, suas capacidades e histórico.

### 5.1 A Arquitetura do classificador

A classificação automática de chamados proposta consiste na forma a reduzir os erros nos processo de classificação, este foi dividido em etapas sucessivas.

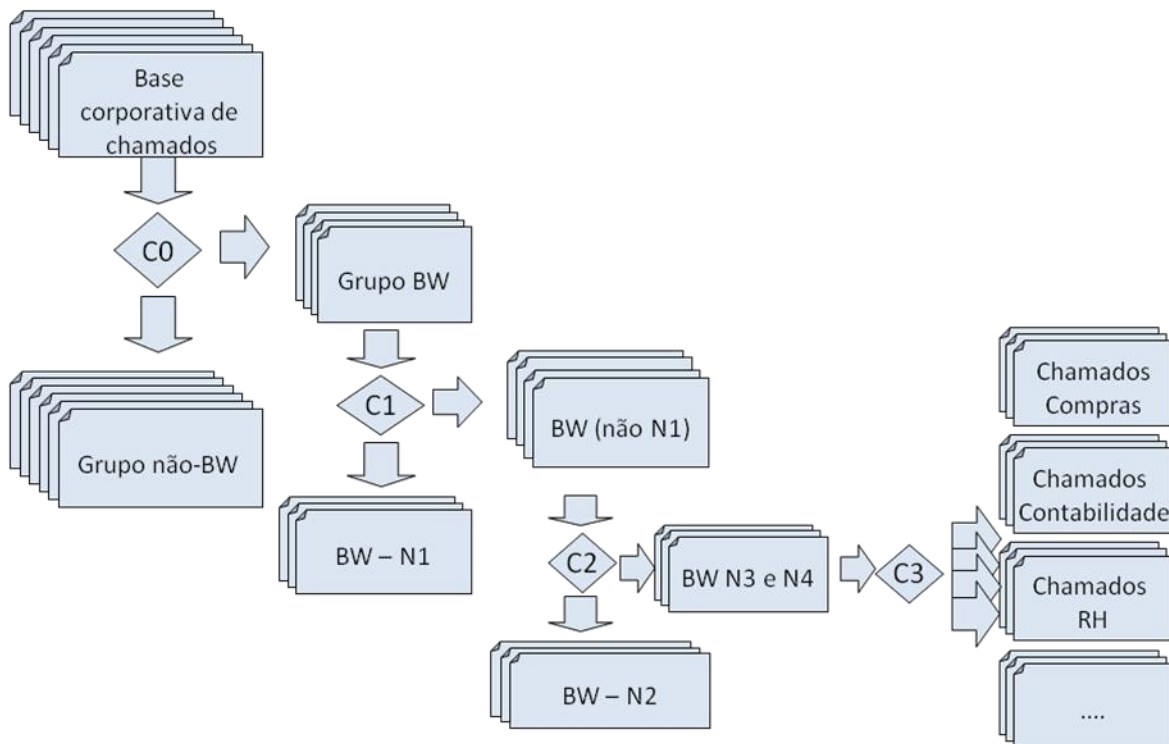
A primeira etapa da classificação, batizada de C0, consistirá na separação entre todos os chamados submetidos em dois grupos: Aqueles que são relevantes para esse estudo (isto é, aqueles cuja ferramenta envolvida é o BW) e os demais, referentes às outras ferramentas suportadas pela organização.

A etapa seguinte, C1, envolverá a divisão de todos os chamados relevantes entre os que devem ser atendidos pelas mesas no primeiro nível de atendimento das demais. Conforme apresentado anteriormente, funcionalmente, todas as mesas do primeiro nível atendem o mesmo tipo de problema, a existência de mesas distintas se deve essencialmente ao fato de que existem diversas formas de entrada de dados. No caso dos chamados de BW, existe uma mesa N1 que atende os chamados abertos pela internet, outra mesa atende os chamados abertos por telefone e uma terceira mesa localizada em Macaé, que atende aos chamados abertos por telefone pelos usuários localizados no norte do estado do Rio. Embora cada uma possua responsabilidade sobre um conjunto distinto de chamados, o tipo de chamado atendido por elas é o mesmo. Isso faz com que os chamados atendidos por mesas do primeiro nível representem todos eles uma mesma classe de chamados.

A terceira etapa do processo (C2) envolve a separação dos chamados relativos ao nível N2 de atendimento. É visada a separação dos chamados técnicos, como problemas de hardware, ou dúvidas dos usuários sobre questões técnicas do sistema das dúvidas referentes aos processos de negócio específicos, a serem tratados pelos níveis N3 e N4.

A etapa final da classificação (C3) visa distribuir os chamados restantes, isso é, aqueles que são referentes a algum dos processos de negócio entre as suas respectivas mesas de atendimento.





*Figura 10 - A arquitetura do classificador implementado*

## 5.2 Seleção de dados

Para a realização do experimento, foram utilizados chamados retirados do banco de dados oficial do Lig-TI. A ferramenta oficial possui uma interface que permite a exportação dos arquivos no formato XML.

Conforme apresentado no primeiro capítulo, uma das premissas deste experimento é que seriam tratados chamados marcados com a solução BW. Foram considerados para o experimento, tanto para os conjuntos de treinamento como para os conjuntos de testes apenas chamados com o status fechado. Isto significa que chamados cancelados e chamados que ainda não foram concluídos estão excluídos do experimento. Adicionalmente, podemos afirmar que a ultima mesa atribuída a um chamado fechado é o

grupo responsável pela sua solução.

A tabela a seguir lista todas as mesas envolvidas no atendimento dos chamados relativos à solução BW, junto com seus respectivos níveis de atendimento e a quantidade de chamados que cada uma delas recebeu no período estudados, o primeiro semestre de 2009.

Nível e mesa de atendimento	Número de Chamados
<b>N1</b>	<b>794</b>
TI-RIO-HD-SAP PETROBRAS	755
TI-RIO-HD-SAP WEB	39
<b>N2</b>	<b>382</b>
TI-RIO-HD-SAP BW	382
<b>N3</b>	<b>1580</b>
TI-RIO-ME-SAP N3 BW CO PS PO	542
TI-RIO-ME-SAP N3 BW FI	234
TI-RIO-ME-SAP N3 BW MM	342
TI-RIO-ME-SAP N3 BW PMLT PPQM	196
TI-RIO-ME-SAP N3 BW SD	195
TI-RIO-ME-SAP N3 PERF ACESSO	71
<b>Total geral</b>	<b>2756</b>

*Tabela 4 – Distribuição de chamados no primeiro semestre de 2009*

Quando um chamado é selecionado para o conjunto de testes sua classificação é usada apenas para a validação do resultado obtido, não afetando de nenhuma forma o funcionamento do classificador.

### 5.3 Pré-Processamento dos textos

Tipicamente, algumas técnicas são aplicadas melhorar os resultados dos processos de classificação envolvendo textos. Esse processo, visa tornar os elementos textuais mais representativos, eliminando elementos que podem gerar ambigüidade ou atrapalhar a classificação. O processo envolve duas etapas:

A primeira coisa a se fazer é a remoção de stopwords dos textos. Stopwords são palavras que podem ser consideradas como não relevantes para a análise de textos. Na maioria das vezes são artigos, preposições e conjunções. Essas palavras podem, falsamente, indicar a semelhança entre dois documentos aonde aparecem, devido a sua natureza muito comum em textos. A tabela abaixo apresenta a lista nas 20 palavras mais comuns no conjunto de dados analisado.

Palavra	Frequência	Importância
de	3302	Não – preposição
BW	2764	Sim
para	2446	Não – preposição
o	2415	Não – artigo
acesso	2086	Sim
a	1848	Não – artigo/preposição
usuário	1765	Não – Palavra comum
não	1614	Não – advérbio
no	1607	Não – preposição
SP	1478	Sim
do	1288	Não – preposição
e	1220	Não – conjunção
que	1125	Não – conjunção
remoto	1124	Não – palavra comum
favor	1040	Não – Palavra comum
solicita	933	Não – palavra comum
ao	921	Não – Artigo e preposição
query	918	Sim
com	779	Não – preposição
senha	753	Sim
usuária	738	Não – Palavra comum

*Tabela 5 – Palavras mais comuns*

Algumas delas foram consideradas, após uma avaliação, como relevantes para a classificação, por exemplo: BW, acesso, senha, SAP e PM (abreviação de Plant Maintenance). Outras palavras foram marcadas como irrelevantes, estão na lista por se tratarem de palavras comuns em chamados, por exemplo: “usuário”, “por favor”, “acesso remoto”, “solicito” e “usuária”. Tais palavras não servem de indicação sobre a classe do chamado, e podem ser acrescentadas a lista de stopwords para serem desconsideradas.

A segunda atividade é o stemming, que consiste no agrupamento de palavras que possuem a mesma raiz morfológica, tal processo ajuda a encontrar semelhanças entre documentos que possam conter palavras com o mesmo radical e sufixos, diferentes, como flexões de número ou gênero. (Porter, 80) descreve um algoritmo de stemming que até hoje é um dos mais usados, inclusive gerando derivações em outras línguas. Analisando a tabela das palavras mais comuns acima, podemos observar que palavras como “usuário” e “usuária” ou “query” e “queries” são apenas flexões de gênero e número, respectivamente, da mesma palavra e podem ser consideradas como igual para fins de classificação. Como exemplo “query” e “queries” igualmente indicam que um chamado pode tratar de banco de dados.

## **5.4 Validação dos resultados**

Para validar os resultados obtidos, todos os chamados analisados já possuem sua classe definida a priori, com isso possuímos uma forma de validar a classificação gerada. Caso ela seja igual à classificação original, consideramos um sucesso, caso contrário, uma falha do classificador.

De forma a garantir que os resultados produzidos no experimento sejam independentes da seleção do conjunto inicial de treinamento, foi adotado o procedimento de classificação cruzada. O conjunto de dados é dividido em  $k$  subconjuntos, chamados de folds. O processo de classificação é executado  $k$  vezes, aonde cada um dos subconjuntos diferentes é usado como conjunto de testes e os  $k-1$  conjuntos restantes sendo usados como base de treinamento. Dessa forma, cada um dos chamados do conjunto é classificado uma vez e usado como base de testes nas demais rodadas.

### **5.5 O Experimento**

No experimento desenhado visa-se testar cada uma das etapas do classificador será testada isoladamente, com um conjunto de dados selecionado para verificar seu funcionamento em condições ótimas. Os resultados do experimento servirão de base para a criação do classificador final, que irá implementar todas as etapas do classificador simulando a aplicação do processo proposto em uma situação mais próxima de um cenário real.

Após a preparação do texto, é iniciado o trabalho de classificação. Serão aplicadas diversas combinações de algoritmos, seleção de atributos do chamado e técnicas de processamento de texto.

Essas classificações envolverão os três algoritmos escolhidos para o trabalho. Os classificadores serão avaliados se possuem melhores resultados com a aplicação de pré processamento nos textos ou não. Será verificada a eficácia de uma classificação usando o texto do chamado, apenas o identificador do usuário e sua lotação e também uma combinação de todos esses elementos.

Durante a execução dessas etapas individuais, serão coletadas informações sobre a

acurácia, a precisão, cobertura e a medida F médias para os conjuntos analisados. A medida de custo de repasse não poderá ser determinada na análise individual, pois não é possível determinar qual é o destino final de um chamado quando ele é incorretamente classificado em uma das etapas.

#### **5.5.1 Resultados experimentais da classificação C0:**

O objetivo dessa classificação é encontrar todos os chamados relativos a uma solução específica. Foi escolhida a solução BW, pelos motivos já citados. Embora tenha sido dito que o escopo desse trabalho se limita aos chamados de uma ferramenta específica, o BW, essa etapa da classificação foi projetada para demonstrar que soluções análogas podem ser adotadas para todas as ferramentas corporativas.

Foram incorporados ao conjunto de chamados listados no início desse capítulo, todos os eles relativos ao BW, um grupo de 1150 chamados referentes a outras soluções suportadas pelo help desk, esses chamados são envolvem vários níveis e soluções diferentes e todos foram criados no mesmo período temporal do outro conjunto, o primeiro semestre de 2009.

As tabelas abaixo apresentam os resultados coletados:

Algoritmo	Atributos	Processamento	Acuracia	Precisão	Recall	F
kNN	Chave e lotação	-	0.806	0.8	0.806	0.794
kNN	Texto	Texto Original	0.888	0.887	0.888	0.887
kNN	Texto	Texto processado	0.936	0.935	0.936	0.955
kNN	Texto, Chave e Lotação	Texto Original	0.892	0.891	0.892	0.891
kNN	Texto, Chave e Lotação	Texto processado	0.94	0.939	0.94	0.939
naive Bayes	Chave e lotação	-	0.801	0.798	0.801	0.782
naive Bayes	Texto	Texto Original	0.903	0.913	0.903	0.905
naive Bayes	Texto	Texto processado	0.938	0.941	0.938	0.939
naive Bayes	Texto, Chave e Lotação	Texto Original	0.919	0.925	0.919	0.92
naive Bayes	Texto, Chave e Lotação	Texto processado	0.944	0.946	0.944	0.984
SVM	Chave e lotação	-	0.806	0.802	0.806	0.789
SVM	Texto	Texto Original	0.961	0.961	0.961	0.961
SVM	Texto	Texto processado	0.966	0.966	0.966	0.966
SVM	Texto, Chave e Lotação	Texto Original	0.963	0.963	0.963	0.963
SVM	Texto, Chave e Lotação	Texto processado	0.97	0.969	0.97	0.969

*Tabela 6- Resultados do experimento com a etapa C0*

Pode-se perceber que os resultados dessa classificação foram muito satisfatórios. Com todos os algoritmos avaliados é perceptível que os textos permitem uma distinção clara entre as diferentes ferramentas. Outra observação pertinente é que usando apenas os atributos que identificam o usuário, a taxa de acerto é significativamente. Isso demonstra que os usuários trabalham focados com ferramentas específicas.

### **5.5.2 Resultados experimentais da classificação C1**

Essa classificação consiste na avaliação se a solução de um determinado chamado

é da competência do primeiro nível de atendimento ou não. Essa classificação simula a avaliação que ocorre quando um chamado é criado em uma mesa N1, aonde o analista precisa determinar se o atendimento deve ser concluído localmente repassado para algum nível superior mais competente. O conjunto a ser testado nesse experimento é o conjunto descrito na tabela 4, com no início desse capítulo, que consiste em 2746 chamados de BW, sendo 794 resolvidos em mesas N1. Os resultados são apresentados a seguir:

Algoritmo	Atributos	Processamento	Acuracia	Precisão	Recall	F
kNN	Chave e lotação	-	0.77	0.758	0.77	0.761
kNN	Texto	Texto Original	0.913	0.924	0.913	0.915
kNN	Texto	Texto processado	0.921	0.923	0.921	0.922
kNN	Texto, Chave e Lotação	Texto Original	0.918	0.927	0.918	0.92
kNN	Texto, Chave e Lotação	Texto processado	0.951	0.951	0.951	0.951
naive Bayes	Chave e lotação	-	0.712	0.507	0.712	0.592
naive Bayes	Texto	Texto Original	0.94	0.941	0.94	0.94
naive Bayes	Texto	Texto processado	0.952	0.952	0.952	0.952
naive Bayes	Texto, Chave e Lotação	Texto Original	0.939	0.941	0.939	0.94
naive Bayes	Texto, Chave e Lotação	Texto processado	0.951	0.951	0.951	0.951
SVM	Chave e lotação	-	0.766	0.751	0.766	0.75
SVM	Texto	Texto Original	0.954	0.954	0.954	0.954
SVM	Texto	Texto processado	0.956	0.957	0.956	0.956
SVM	Texto, Chave e Lotação	Texto Original	0.955	0.955	0.955	0.955
SVM	Texto, Chave e Lotação	Texto processado	0.968	0.968	0.968	0.968

*Tabela 7- Resultados do experimento com a etapa C1*



A Classificação C1 assim como a C0, apresentou taxas muito boas de acertos. Ao contrario do nível anterior, pode-se ver que a adição do usuário e de sua lotação aos atributos analisados não melhorou significativamente os resultados. Isso se deve ao fato de que os tipos de solicitações atendidas pelo N1 são as mais básicas, como desbloqueio de senha, solicitação de acesso ou esclarecimentos sobre down-time do sistema. Tais solicitações são muito comuns e feitas por todos os usuários eventualmente.

### **5.5.3 Resultados experimentais da classificação C2**

Essa classificação funciona de forma análoga a C1. Simula a avaliação realizada pelos técnicos das mesas N2 e consiste na determinação de se o problema deve ser solucionado pela mesa ou repassado para o próximo nível para atendimento da equipe especializada. Para esse teste são usados os mesmos dados da etapa anterior, menos os chamados pertinentes ao N1, que já deveria ter sido atribuídos às suas mesas. Restam 1580 chamados de N3 382 chamados de N2 que devem ser encontrados nessa classificação. Os resultados estão apresentados a seguir:

Algoritmo	Atributos	Processamento	Acurácia	Precisão	Recall	F
kNN	Chave e lotação	-	0.826	0.807	0.826	0.811
kNN	Texto	Texto Original	0.728	0.784	0.728	0.748
kNN	Texto	Texto processado	0.709	0.77	0.709	0.731
kNN	Texto, Chave e Lotação	Texto Original	0.743	0.786	0.743	0.759
kNN	Texto, Chave e Lotação	Texto processado	0.751	0.786	0.751	0.765
naive Bayes	Chave e lotação	-	0.805	0.649	0.805	0.719
naive Bayes	Texto	Texto Original	0.774	0.853	0.774	0.794
naive Bayes	Texto	Texto processado	0.802	0.853	0.802	0.817
naive Bayes	Texto, Chave e Lotação	Texto Original	0.775	0.854	0.775	0.795
naive Bayes	Texto, Chave e Lotação	Texto processado	0.799	0.853	0.799	0.814
SVM	Chave e lotação	-	0.83	0.809	0.83	0.808
SVM	Texto	Texto Original	0.826	0.828	0.826	0.827
SVM	Texto	Texto processado	0.816	0.822	0.816	0.819
SVM	Texto, Chave e Lotação	Texto Original	0.805	0.649	0.805	0.719
SVM	Texto, Chave e Lotação	Texto processado	0.83	0.828	0.83	0.829

*Tabela 8- Resultados do experimento com a etapa C2*

Os resultados são dessa etapa não apresentaram significativas diferenças entre as combinações distintas de classificador, mesmo a adição da identificação do usuário e a lotação não causam melhorias significativa nos resultados. De qualquer forma, os resultados permanecem em um intervalo satisfatório.

### 5.5.4 Resultados experimentais da classificação C3

Essa classificação tem como objetivo separar os chamados que já foram classificados como pertinentes ao N3 entre todas as mesas desse nível que existem para a solução em questão. Foram utilizados todos os 1580 chamados de N3 registrados no primeiro semestre de 2009 nesse experimentos. Esses chamados compõem 6 meses de atendimento, e devido a esse aumento no número de classes espera-se que os indicadores dos classificadores apresentem alguma queda. Os valores obtidos foram os seguintes:

Algoritmo	Atributos	Processamento	Acurácia	Precisão	Recall	F
kNN	Chave e lotação	-	0.716	0.715	0.716	0.714
kNN	Texto	Texto Original	0.588	0.595	0.588	0.585
kNN	Texto	Texto processado	0.587	0.598	0.587	0.589
kNN	Texto, Chave e Lotação	Texto Original	0.639	0.648	0.639	0.638
kNN	Texto, Chave e Lotação	Texto processado	0.639	0.648	0.639	0.638
naive Bayes	Chave e lotação	-	0.645	0.712	0.645	0.636
naive Bayes	Texto	Texto Original	0.709	0.755	0.709	0.715
naive Bayes	Texto	Texto processado	0.763	0.789	0.763	0.767
naive Bayes	Texto, Chave e Lotação	Texto Original	0.714	0.764	0.714	0.72
naive Bayes	Texto, Chave e Lotação	Texto processado	0.777	0.803	0.777	0.78
SVM	Chave e lotação	-	0.779	0.793	0.779	0.775
SVM	Texto	Texto Original	0.751	0.753	0.751	0.751
SVM	Texto	Texto processado	0.753	0.756	0.753	0.752
SVM	Texto, Chave e Lotação	Texto Original	0.83	0.833	0.83	0.829
SVM	Texto, Chave e Lotação	Texto processado	0.836	0.839	0.836	0.836

*Tabela 9- Resultados do experimento com a etapa C3*

Os resultados da classificação C3 demonstram claramente a vantagem tanto da escolha do algoritmo SVM como na importância de se usar a identificação do usuário e a lotação. Como a divisão entre as mesas N3 segue a lógica dos processos de negócio e conseqüentemente a estrutura da empresa, é de se esperar que o uso de informações sobre o usuário e sua lotação pudessem influenciar positivamente a classificação.

## **5.6 A Implementação**

Uma vez que todas as etapas foram testadas individualmente e apresentam resultados satisfatórios, o próximo passo é a verificação do processo proposto através da implementação do classificador.

A aplicação proposta consiste no uso sucessivo das etapas de classificações C0, C1, C2 e C3. Cada uma delas recebe de entrada um conjunto de dados que é o resultado da etapa anterior. Esse seqüenciamento faz com que chamados incorretamente classificados em um nível prejudiquem a classificação dos níveis subseqüentes, propagando os erros e aproximando mais a simulação de uma situação cotidiana.

Conforme pode ser observados nos resultados dos experimentos com as etapas individuais, os indicadores de qualidade de cada classificador variam de acordo com o nível da classificação, algoritmo escolhido e os atributos escolhidos. Embora seja possível detectar alguns padrões recorrentes em todas as etapas, como o desempenho inferior do classificador kNN em relação aos demais, ou a importância do processamento de texto, não é possível perceber uma estratégia única que possui o melhor desempenho e todas as situações. Uma forma produzir um resultado ótimo é testar combinações diferentes para cada uma das etapas, maximizando o número de erros, ou o custo de

repasses do resultado final.

Como optou-se por não desprezar nenhuma das combinações de algoritmo e seleção de atributos na aplicação do classificador, com a intenção de buscar um resultado mais completo, o número de combinações é consideravelmente alto: três algoritmos e três combinações de atributos perfazem 9 combinações por etapa de classificação. Este número, elevado a quarta potência, para cada um dos níveis no experimento, produz uma quantidade altíssima de combinações a serem testadas. Para reduzir esse número foi aplicado o algoritmo de otimização Hill Climbing. Conforme discutido na revisão, esse algoritmo é simples e considerado um bom ponto de partida para testes com otimização.

A aplicação da otimização nessa situação é interessante pois, ao se iniciar a classificação de um novo nível, deve-se procurar a melhor combinação de algoritmos e seleções para a classificação do nível atual, isto é, a classificação que minimiza os erros nos chamados repassados para o nível seguinte. Dessa forma, um conjunto de saída avaliado com pior é descartado e evita-se o trabalho de classificação do nível seguinte com os seus dados.

### **5.6.1 Resultados obtidos**

A primeira etapa do classificador sequencial é a classificação C0, que segrega todos os chamados entre os referentes ao sistema BW e os demais. Conforme discutido anteriormente, este trabalho se limita aos chamados de BW, portanto não foram adicionados a esse conjunto de treinamento chamados de outras soluções.

Classificação C0 BW e não BW			
	usuário e lotação	texto	usuário, lotação e texto
kNN	F=1 C=0	F=1 C=0	F=1 C=0
n Bayes	F=1 C=0	F=1 C=0	F=1 C=0
SVM	F=1 C=0	F=1 C=0	F=1 C=0

*Tabela 10 - Resultados da etapa C0 da classificação sequencial*

Os resultados, conforme era esperado, foram excelentes para todas as combinações, porém não possuem nenhuma representatividade prática. Todos os 2757 chamados marcados como BW foram corretamente repassados para a próxima etapa.

A Classificação C1 consiste no processamento de todos os chamados de BW para segregação dos chamados N1. A tabela a seguinte apresenta os valores de F e o custo de repasses para as combinações avaliadas:

Classificação C1: (N1 e não N1)			
	usuário e lotação	texto	usuário, lotação e texto
kNN	F=0.667 C=385	F=0.753 C= 70	F=0.762 C=165
n Bayes	F=0.419 C=784	F=0.826 C=104	F=0.823 C=106
SVM	F=0.704 C=213	<b>F=0.828 C= 64</b>	F=0.844 C= 64

*Tabela 11 - Resultados da etapa C1 da classificação sequencial*

Os resultados são semelhantes aos valores obtidos na classificação C1 individual. Os menores custos encontrados foram na aplicação do algoritmo SVM. Conforme definido, o custo foi calculado com base nos chamados que foram incorretamente classificados como de N1. Chamados N3 foram computados com peso dois, pois exigirão dois repasses até que sejam atribuídos a sua mesa correta, e chamados de N2 foram computados com peso unitário, pois seu atendimento exigirá somente um repasse.

A próxima etapa é a classificação C2, aplicando todos os 1973 chamados que não

foram classificados como N1 na etapa anterior. Os valores do custo de repasses e da medida F são apresentados a seguir:

Classificação C2 (N2 e não N2)			
	usuário e lotação	texto	usuário, lotação e texto
kNN	F=0.787 C=223	F=0.725 C=208	F=0.733 C=303
n Bayes	F=0.699 C=348	F=0.782 C=109	<b>F=0.778 C=106</b>
SVM	F=0.779 C=241	F=0.788 C=177	F=0.78 C=212

*Tabela 12- Resultados da etapa C2 da classificação sequencial*

Os resultados mostram que o algoritmo probabilístico de naive Bayes possui um desempenho bastante superior aos demais. O custo de repasses, conforme definido, foi calculado como o número de chamados incorretamente classificados como N2, tendo em vista que o esforço para devolver um chamado para o N1 e para repassar um chamado para o N3 correspondente é o mesmo.

A última etapa é classificação C3, que consiste na divisão de todos os chamados que não foram classificados como N1 e N2 nas mesas de atendimento do nível N3, que totalizam 1400 chamados.

Classificação C3: Mesas de Atendimento			
	usuário e lotação	texto	usuário, lotação e texto
kNN	F=0.693 C= 808	F=0.642 C=1024	F=0.703 C=842
n Bayes	F=0.563 C=1170	F=0.759 C= 730	F=0.767 C=714
SVM	F=0.699 C= 798	F=0.775 C= 632	<b>F=0.797 C=558</b>

*Tabela 13- Resultados da etapa C3 da classificação sequencial*

O melhor resultado obtido, de acordo com o custo dos repasses é a aplicação do algoritmo SVM para classificação dos textos, nome e lotação do usuário. O custo dos repasses é calculado da seguinte forma: Chamados de níveis anteriores que foram incorretamente classificados como N3 devem ser devolvidos, sendo que o custo é

dobrado para chamados de N1, pois eles não devem ser devolvidos diretamente e devem passar pelo N2. Chamados do próprio nível N3 atribuídos a uma mesa incorreta são repassados também com o custo de dois. Os analistas N3 que recebem um chamado incorreto devem devolver o chamado para o nível anterior, só em seguida eles podem ser atribuídos à mesa correta.

### **5.6.2 Ganhos Percebidos**

Se aplicarmos a medição do custo de repasses utilizada no processo atual, podemos facilmente calcular um valor que indica a quantidade de repasses que devem ser feitos para o atendimento dos chamados no processo atual. Como todos os chamados são inicialmente atribuídos ao N1 para subsequente repasse, no processo manual, podemos considerar que todos os chamados que são atendidos nas mesas N1 possuem custo 0, enquanto que todos os chamados atendidos por mesas N2 recebem um repasse (custo 1) e todos os chamados atendidos pelo N3 sofrem dois repasses (custo 2). Dessa forma, o custo total de designação de chamados no sistema atual pode ser calculado como:

$$\text{Custo total de repasses} = (794 \times 0) + (382 \times 1) + (1580 \times 2) = 3546$$

Em um caso de classificação ideal, aonde todos os chamados pudessem ser corretamente classificados, o custo de repasses seria igual a zero, eliminando efetivamente o tráfego de chamados entre as mesas de atendimento visando a classificação correta de um chamado. Nesse caso, a eliminação do tempo de designação dos chamados acarretaria num tempo total de atendimento de chamados reduzido em 15%.

O custo total de repasses obtido no experimento é de 558. Isso indica que a aplicação do processo de classificação automático da forma apresentada nesse trabalho



tem a possibilidade de reduzir o volume de repasses em 85% do tempo empregado atualmente. Essa redução acarretaria num tempo total de atendimento reduzido em 12,75% em média para todos os chamados referentes a solução BW.

## Capítulo 6 – Conclusões

Após a exposição da proposta de trabalho, dos conceitos envolvidos e dos resultados obtidos com a aplicação desenvolvida, esse capítulo apresenta as conclusões finais deste trabalho.

Foi proposto que, de forma a reduzir o tempo necessário para tratamento de problema de TI, a aplicação de técnicas de classificação de textos poderia melhorar o tempo gasto na atividade de encaminhamento dos chamados para o grupo mais indicado para seu atendimento. Nessa abordagem, os grupos de atendimento são vistos como categorias, e os relatórios de chamados cadastrados no sistema são os documentos atribuídos a essas categorias. Palavras que compõe os textos são os atributos dos documentos devem ser usados para sua classificação. Um novo processo de atendimento, que consiste em uma versão revisada do processo atual, foi apresentada para mostrar como a aplicação da classificação poderia facilitar o processo.

Um experimento foi realizado conforme as condições descritas e os resultados apresentados de forma tabular. Conforme relatado, a aplicação das técnicas de classificação de textos nos chamados, nas condições estudadas, possibilita que os chamados possam ser direcionados para seus grupos responsáveis de forma mais rápida do que o processo manual, que é utilizando nas estruturas de help desk das organizações e empresas hoje.

A aplicação desenvolvida aplicou o processo proposto num conjunto de chamados reais e os resultados obtidos indicam que, mesmo com uma taxa de erros razoável, os ganhos no tempo de atendimento são significativos.

O processo proposto apresentou resultados promissores, indicando que os textos são uma boa forma de encontrar semelhanças entre os chamados. Essa relação possibilita que outras informações sejam extraídas de uma base e atribuídas a um novo chamado. Essas informações podem incluir medidas de impacto que um problema pode causar, tanto em termos de indisponibilidade de sistema como em problemas nas regras de negócio, isso possibilitaria uma priorização especial no atendimento.

Outra idéia promissora seria encontrar propostas de soluções para o problema relatado em um chamado, consultando informações sobre como outros chamados semelhantes foram encerrados. O classificador poderia gerar uma lista de sugestões ou recomendações e propor para os analistas.

Foi percebido que diferentes técnicas de classificação, em combinação com diferentes seleções de atributos para classificação de cada nível podem obter resultados melhores. Dentre as alternativas testadas foi identificada qual a melhor combinação para cada nível da classificação seqüencial, juntamente com uma breve análise sobre os motivos que podem ter levado a combinação a ter os melhores resultados.

Por fim, conclui-se que os resultados desse trabalho habilitam os interessados em tornar os processos de atendimento mais ágeis, reduzindo os tempos de interrupção de serviços quem podem ser críticos para as organizações, reduzindo potenciais prejuízos para os negócios e aumentando a eficiência operacional.

## Referências Bibliográficas

1. ANVIK, J, HIEW, L. MURPHY, G. C., Who should fix this bug? in ICSE 06: Proceeding of the 28th international conference on Software engineering, pp. 361-370 (New York, NY, USA: ACM Press, 2006).
2. BARBOSA, Rodolfo e ALVES, Aluísio. Petrobras ultrapassa Microsoft em valor de mercado, diz estudo. São Paulo: Agência Reuters, 19/05/2008
3. BOSER. B, GUYON, I. e VAPNIK, V. - A Training Algorithm for Optimal Margin Classifiers, Proc. of the 5th Annual ACM Workshop on Computational Learning Theory, pp.144-152, ACM Press, 1992
4. CAMARGO, Y. B. L. Abordagem Lingüística na Classificação de Textos em Português. Dissertação de M.Sc. COPPE/UFRJ, Rio de Janeiro, RJ 2007
5. CARBONELL JG, MICHALSKI, RS; MITCHELL TM - Machine learning: An artificial intelligence approach, Morgan Kaufmann 1986. ISBN 0934613001
6. CHAKRABARTI S. Mining the Web: Discovering Knowledge from Hyper-text Data, Morgan Kaufmann, 2002
7. COVER, T., HART, P., 1967, "Nearest neighbor pattern classification", Information Theory, IEEE Transactions on, v. 13, n. 1, pp. 21-27.

8. CUBRANIC, D. e MURPHY, G.C. – Automatic bug triage using text classification. In Proceedings of Software Engineering and Knowledge Engineering, pages 92-97, 2004.
9. DUMAIS, S., PLATT, J., HECKERMAN, D., e SAHAMI, M. 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of the Seventh international Conference on information and Knowledge Management (Bethesda, Maryland, United States, November 02 - 07, 1998). CIKM '98. ACM, New York, NY, 148-155.
10. EADS, D.R; HILL, D.; DAVIES, S.; PERKINS, S.J.;MA, J; PORTER, R.B. ; THEILER, J.P. - Genetic algorithms and support vector machines for time series classification, in: B. Bosacchi, D.B. Fogel, J.C. Bezdek (Eds.), Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation V, Proceedings of the SPIE, vol. 4787, 2002, pp. 74–85.
11. EIBE, I. W.; WITTEN, I. H; FRANK, E.; TRIGG, L.; HALL, M.; HOLMES, G.; CUNNINGHAM, S.J.; Weka: Practical Machine Learning Tools and Techniques with Java Implementations. In:Proc ICONIP/ ANZIIS/ANNES99 Future Directions for Intelligent Systems and Information Sciences 1999.

12. FRIEDMAN, J.H. Flexible Metric Nearest Neighbour Classification - technical report, Stanford Univ., Nov. 1994.
13. FRIEDRICH, F; IGEL C; Evolutionary tuning of multiple SVM parameters. Trends in Neurocomputing: 12th European Symposium on Artificial Neural Networks 2004. 2005
14. FURTADO, M. I. V., Inteligência Competitiva para o Ensino Superior Privado: uma Abordagem através da Mineração de Textos, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004.
15. GARNER, S.R. 1995. WEKA: The Waikato Environment for Knowledge Analysis. In Proceedings of the New Zealand Computer Science Research Students Conference, pages 57–64.
16. HAN, J. e KAMBER, M - DataMining: Concepts and Techniques, Morgan Kaufmann, San Francisco, CA, 2001.
17. HASTIE, T e TIBSHIRANI, R - Discriminant Adaptive Nearest Neighbor Classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 6, pp. 607-616, June, 1996.

18. JENSEN, V. FINN. Bayesian Networks and Decision Graphs. Springer-Verlag. 2001.
19. JOACHIMS, T. Text Categorization with support vector machines: learning with many relevant features. In: Proceedings of the 10th European Conference on Software Engineering, p137-142. 1998
20. LEWIS, D., 1998, "Naive Bayes at forty: The independence assumption in information retrieval", pp. 4-15, Springer Verlag, Heidelberg, DE.
21. LOPES, M. C. S., Mineração de dados textuais utilizando técnicas de clustering, para o idioma português, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2004.
22. LOUREIRO, S. M., MARGOTO, L. R., VAREJÃO, F. M., et al., 2005, "Um mecanismo automático para busca de parâmetros de técnicas de classificação utilizando algoritmos genéticos". XXV Simpósio Brasileiro de Computação - SBC, V Encontro Nacional de Inteligência Artificial – ENIA
23. NEVES, M. L., 2001. PubsFinder - um Agente Inteligente para Busca e Classificação de Páginas de Publicações. Dissertação de M.Sc. Centro de Informática da UFPE, Recife.

24. PINHEIRO, A. C. L. - Ambiente Analítico e Qualidade de Informações: um estudo de caso da implantação do SAP BW na Petrobras. Dissertação de Mestrado. Pontifícia Universidade Católica do Rio de Janeiro. 2006.
25. PORTER, M.F. An algorithm for suffix stripping, Program 14 (1980)
26. REDMAN, T. The impact of poor data quality on the typical enterprise - Communications of the ACM, Volume 41, Issue 2 (February 1998)
27. RIZZI, C. B., WIVES, L. K., OLIVEIRA, J. P. M., ENGEL, P.M. , 2000, "Fazendo uso da Categorização de Textos em Atividades Empresariais", International Symposium on Knowledge Management/Document Management - ISKDM/DM 2000, pp. 251-268, Curitiba, Brasil.
28. ROBERTSON, S. E., 1997, "The probability ranking principle in IR", pp. 281-286.<http://portal.acm.org/citation.cfm?id=275701>
29. RUSSELL, S; NORVIG - Artificial Intelligence: A Modern Approach. Prentice Hall, 1995.



30. SEBASTINANI, F. Machine learning in automated text categorization. *ACM Comput. Surv.* 34, 1 (Mar. 2002), 1-47.
31. SILVA, C. F., FAGUNDES, R., 2007, "Categorização de Textos da Língua Portuguesa com Árvores de Decisão, SVM e Informações Lingüísticas". XXVII Simpósio da Sociedade Brasileira de Computação, Rio de Janeiro, Brasil
32. WEILL, P., SUBRAMANI, M.R., e BROADBENT, M. "Building IT Infrastructure for Strategic Agility," *Sloan Management Review* (44:1) 2002, pp 57-65.
33. WEISS, C, PREMRAJ, R, ZIMMERMANN, T. ZELLER, A. How Long Will It Take to Fix This Bug? in *MSR 07: Proceedings of the Fourth International Workshop on Mining Software Repositories* (Washington, DC, USA: IEEE Computer Society, 2007).
34. YANG, Y. and LIU, X. A re-examination of text categorization methods. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

35. ZHANG, H., BERG, A. C., MAIRE, M e MALIK, J. - SVM-KNN:

Discriminative Nearest Neighbor Classification for Visual Category Recognition.

In Proceedings of the 2006 IEEE Computer Society Conference on Computer

Vision and Pattern Recognition - Volume 2 (June 17 - 22, 2006). CVPR. IEEE

Computer Society, Washington, DC, USA