UNIVERSIDADE DE BRASÍLIA
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, CONTABILIDADE E CIÊNCIA DA INFORMAÇÃO E DOCUMENTAÇÃO – FACE
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO E DOCUMENTAÇÃO – CID

DESCOBERTA DE CONHECIMENTO EM TEXTO
APLICADA A UM SISTEMA DE ATENDIMENTO AO CONSUMIDOR

JOSÉ MARCELO SCHIESSL

BRASÍLIA
2007

DESCOBERTA DE CONHECIMENTO EM TEXTO
APLICADA A UM SISTEMA DE ATENDIMENTO AO CONSUMIDOR

JOSÉ MARCELO SCHIESSL

Dissertação apresentada à banca examinadora como requisito parcial à obtenção do Título de Mestre em Ciência Da Informação pelo Programa de Pós-Graduação em Ciência da Informação do Departamento de Ciência da Informação e Documentação da Universidade de Brasília.

Orientadora: Doutora Marisa Bräscher Basílio Medeiros

BRASÍLIA
2007

Universidade de Brasília (UnB)
Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação (FACE)
Departamento de Ciência da Informação e Documentação (CID)
Programa de Pós-Graduação em Ciência da Informação (PPGCINF)

FOLHA DE APROVAÇÃO

Titulo: *"Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor"*

Autor: José Marcelo Schiessl

Área de concentração: Transferência da Informação

Linha de Pesquisa: Arquitetura da Informação

Dissertação submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação e Documentação, da Universidade de Brasília como requisito parcial para obtenção do título de **Mestre** em Ciência da Informação.

Dissertação aprovada em: 12 de abril de 2007

Aprovado por:

Profª Drª. Marisa Bräscher Basílio Medeiros
Presidente – Orientador (UnB/PPGCInf)

Prof. Dr. Jaime Robredo
Membro Interno – (UnB/PPGCInf)

Prof. Dr. Miguel Filho Ferreira de Oliveira
Membro Externo – (CEF)

Prof. Dr. Ailton Gonçalves Feitosa
Suplente – (IESB)

À Solange, Ingrid e Karin.

AGRADECIMENTOS

À Solange, minha esposa, e às minhas filhas Ingrid e Karin pelo porto seguro que têm proporcionado a minha vida;

À professora Marisa pela dedicação e paciência na orientação deste projeto;

Ao professor Miguel pelo incentivo e energia que sempre permearam as nossas conversas;

Ao professor Robredo pela disponibilidade e orientações precisas;

Ao professor Ailton pela amizade de mais de duas décadas;

À Ony que, com seu carinho e compreensão, foi fundamental para que eu tivesse disponibilidade para fazer deste projeto uma realidade;

À Solange Cassis e ao Marcos que doaram tempo e sabedoria para que eu pudesse ter acesso à matéria-prima, os dados, e às pessoas que me ensinaram o funcionamento do SAC;

Aos amigos que, mesmo neste período de mergulho intelectual solitário, não desistiram de mim;

A todos que, de alguma forma, me incentivaram e me apoiaram no caminho até aqui.

“O mais importante na comunicação
é ouvir o que não foi dito”.
Peter F. Drucker

RESUMO

Analisa um Serviço de Atendimento ao Consumidor de uma instituição financeira que centraliza, em forma textual, os questionamentos, as reclamações, os elogios e as sugestões, verbais ou escritas, de clientes. Discute a complexidade da informação armazenada em linguagem natural para esse tipo de sistema. Visa apresentar alternativa para extração de conhecimento de bases textuais com a criação de agrupamentos e modelo de classificação automática de textos para agilizar a tarefa realizada atualmente por pessoas. Apresenta uma revisão de literatura que mostra a Descoberta de Conhecimento em Texto como uma extensão da Descoberta de Conhecimento em Dados que utiliza técnicas do Processamento de Linguagem Natural para adequar o texto a um formato apropriado para a mineração de dados e destaca a importância do processo dentro da Ciência da Informação. Aplica a Descoberta de Conhecimento em Texto em uma base do Serviço de Atendimento ao Cliente com objetivo de criar automaticamente agrupamentos de documentos para posterior criação de um modelo categorizador automático dos novos documentos recebidos diariamente. Essas etapas contam com a validação de especialistas de domínio que atestam a qualidade dos agrupamentos e do modelo. Cria indicadores de desempenho que avaliam o grau de satisfação do cliente em relação aos produtos e serviços oferecidos para fornecer subsídio à gestão na política de atendimento.

Palavras-chave: Descoberta de Conhecimento em Texto, Mineração de Textos, Mineração de Dados, Descoberta de Conhecimento em Dados.

ABSTRACT

It analyses a Help Desk System of a federal institution that centralizes customer answers, complains, compliments, and suggestions, spoken or written. It argues about information complexity stored in natural language. It intends to present an alternative for knowledge extraction from textual databases by creating clusters and automatic classification model of texts in order to improve the current tasks made by employees. It presents a literature revision that shows the Knowledge Discovery in Text as an extension of Knowledge Discovery in Data that utilizes the Natural Language Processing in order to adequate the text into an appropriated format to data mining and enhances the importance of the process in the Information Science field. It applies the Knowledge Discovery in Text techniques in the Help Desk Database in order to create cluster of documents and, after that, to build an automatic classification model to new documents received every day. These steps need to be validated by specialist in the area to verify the model and clusters quality. It creates performance indexes in order to measure the customer satisfaction related to products and services to provide information for decision makers.

Key Words: Knowledge Discovery in Text, Text Mining, Help Desk System, Data Mining, Knowledge Discovery in Data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

DCT	Descoberta de Conhecimento em Textos;
DCD	Descoberta de Conhecimento em Dados;
SAC	Serviço de Atendimento ao Consumidor;
PLN	Processamento de Linguagem Natural;
KDT	*Knowledge Discovery in Texts;*
KDD	*Knowledge Discovery in Data*;
DM	*Data Mining;*
TM	*Text Mining;*
CI	Ciência da Informação;
RI	Recuperação da Informação;
CC	Ciência da Computação;
GF/IDF	*Global Frequency/Inverse Document Frequency;*
IDF	*Inverse Document Frequency;*

SUMÁRIO

INTRODUÇÃO

O trabalho aplica técnicas de mineração de texto numa base de Serviço de Atendimento ao Consumidor a fim de demonstrar a utilidade da Descoberta de Conhecimento em Textos que consistirá na criação de agrupamentos de textos a partir da coleção de documentos existentes.

Atualmente, os recursos computacionais promovem o acesso à informação de maneira rápida e eficiente, desde que esteja organizada em bancos de dados apropriados à manipulação por computadores. Grande parte da informação eletrônica encontra-se disponível em bases de dados freqüentemente chamadas de não-estruturadas, ou seja, bases de documentos textuais, cujo formato está adequado ao homem que, através da leitura, é capaz de decodificar a informação contida no texto e apreendê-la.

Por outro lado, a quantidade desses documentos produzidos não é passível de ser absorvida pelo homem por esse processo e, dessa maneira, a máquina desempenha um papel fundamental na gestão da informação. Para tal, é necessário o processamento prévio do texto com a finalidade de decodificá-lo e ajustá-lo às estruturas reconhecidas pelos computadores.

A Descoberta de Conhecimento em Texto (DCT) propõe soluções para tratar a informação eletrônica textual com o auxílio de máquinas, visando diminuir o impacto da sobrecarga de informação.

Um Serviço de Atendimento ao Consumidor (SAC) é um serviço que estabelece um canal de comunicação do cliente com a empresa que visa identificar as suas necessidades e seus desejos em relação aos produtos e serviços oferecidos. Uma base de dados de um SAC é uma fonte muito rica de informações que são passadas pelos clientes sem que se faça pesquisa de mercado ou similares que, em geral, são muito caras. O cliente expressa seu ponto de vista espontaneamente motivado pelo desejo de que seja reparado algum dano que a instituição tenha lhe causado, pela vontade de que se implemente alguma sugestão ou, até mesmo, pelo anseio de tecer elogios sobre produtos, serviços ou sobre a própria organização. Geralmente, essas bases de dados contêm a transcrição em linguagem natural da argumentação do cliente.

Esse trabalho aplica a DCT em uma base de SAC de uma instituição financeira com o intuito de mostrar que, com a utilização de ferramentas e metodologias adequadas, é possível maximizar a descoberta e a utilização de informações novas e úteis ainda não identificadas.

Pretende-se destacar a relação entre Descoberta de Conhecimento em Dados (DCD) e a DCT. Para tanto, são apresentadas as visões de autores sobre a DCD, bem como o seu processo de execução. Aborda-se o Processamento de Linguagem Natural (PLN) e sua ligação com a DCD e a argumentação sobre o tema central, a DCT, seus aspectos relevantes e sua fundamentação em outras áreas do conhecimento, além de um esboço de suas aplicações.

1 DEFINIÇÃO DO PROBLEMA

Num Serviço de Atendimento ao Consumidor há uma base de dados textuais na qual são registrados os elogios, reclamações ou sugestões em linguagem natural. Cada registro corresponde a um texto contendo as transcrições de clientes da instituição financeira. Esses textos são fontes de informação importante para gestão da empresa, porém, na forma textual, há uma tarefa árdua de leitura por parte dos analistas para apreender o conteúdo de cada um.

Uma característica interessante é que, dada a quantidade de registros e de analistas diferentes, as relações entre os textos, isto é, as associações entre os temas descritos nos registros não são observadas. Essas informações implícitas, que existem apenas no contexto da análise de vários documentos concomitantemente, não são visualizadas por falta de ferramental apropriado.

Dessa forma, o trabalho manual de leitura, de classificação, de envio de mensagens aos gestores e de respostas aos clientes demanda tanto recursos humanos quanto tempo e, assim, impacta diretamente na velocidade e na qualidade de atendimento esperado pelo cliente.

Como mencionado, o trabalho é feito manualmente e, dado o seu volume, o acúmulo é inevitável. Eventualmente, são executados mutirões de leitura com objetivo de dar vazão às mensagens represadas e de compreender problemas específicos apontados nos seus conteúdos.

Outra característica advinda da manipulação das informações por pessoas é que a classificação das mensagens e o encaminhamento para destinatários corretos estão correlacionados à experiência do profissional, isto é, a chance de erros aumenta proporcionalmente a sua inabilidade em reconhecer o assunto das mensagens e vinculá-las às áreas gestoras. Dessa forma, há que se fazer uma verificação rotineira para reclassificar essas mensagens.

Os agrupamentos criados para categorizar os textos devem passar por revisões, em conseqüência da dinâmica do fluxo de trabalho, com objetivo de identificar novas categorias ou eliminar antigas que não mais se aplicam.

Por último, considera-se que a informação contida nos documentos do SAC é importante para a estratégia de atendimento e, uma vez estruturada, pode nortear a gestão estratégica na definição de políticas para produtos e serviços da organização com o objetivo de melhor atender à demanda do cliente.

2 QUESTÃO DA PESQUISA

É possível, com a aplicação da DCT, gerar indicador relativo à satisfação dos clientes de produtos e serviços de uma instituição financeira que possam orientar a definição de estratégias e políticas de atendimento?

3 OBJETIVOS

3.1 Objetivo Geral

Gerar indicador relativo à satisfação dos clientes de produtos e serviços de uma instituição financeira, a partir da aplicação de DCT em uma base de SAC, visando subsidiar políticas e estratégias de atendimento.

3.2 Objetivos Específicos

- Extrair conhecimento da base SAC com aplicação da DCT e identificar e/ou propor categorias de agrupamento de tipos de reclamações com base no conteúdo das reclamações;

- Criar modelo capaz de classificar automaticamente novos documentos com indicação de assertividade;
- Analisar se a classificação automática obtida por meio da DCT possibilita otimização da classificação humana atualmente adotada;
- Gerar indicador relativo à satisfação dos clientes para subsidiar políticas e estratégias de atendimento.

4 JUSTIFICATIVA

Com os avanços tecnológicos, a utilização de computadores para a manipulação de informação tem substituído a abordagem dispensada ao suporte físico tradicional, o papel, pelo meio eletrônico. Nesse contexto, à medida que os recursos computacionais tornam-se mais confiáveis e acessíveis, o crescimento contínuo dos volumes de dados e a velocidade com que são disseminados contribuem para que a sua administração não seja tarefa trivial.

Com o fenômeno da explosão informacional, o foco de profissionais, que se dedicavam à produção de produtos tangíveis, tem migrado para o tratamento de mensagens e signos (MIRANDA, 2003). Dessa forma, a informação adquire valor de um bem, um patrimônio que pode ser comercializado como qualquer mercadoria.

Além da visão da informação como capital precioso equivalente aos recursos de produção, materiais e financeiros, ela é também vista, por muitas empresas, como elemento estruturante e instrumento de gestão. Assim, a gestão efetiva de uma organização requer a percepção objetiva e precisa dos valores da informação e do sistema de informação (MORESI, 2000).

No contexto organizacional, as funções da administração – planejamento, organização, liderança e controle – são exigências para o sucesso no desempenho da empresa. Nesse sentido, a informação é apoio fundamental a essas funções, principalmente no planejamento e no controle, pois a partir da sua utilização adequada, na hora certa e com a qualidade pretendida, os administradores podem monitorar o progresso de seus objetivos. O insumo de informação com qualidade favorece a tomada de decisão com o propósito de transformar os planos organizacionais em realidade (STONER; FREEMAN, 1995).

Muito tem se falado a respeito da explosão informacional e, freqüentemente, alardeia-se que a informação criada diariamente vem aumentando à medida que novas tecnologias são implementadas e muito pouco destes dados poderão ser percebidos por humanos (LYMAN;VARIAN, 2003).

Além disso, estima-se que 80% dos dados das organizações não podem ser tratados com ferramentas padrão de mineração de dados, pois se encontram em formato textual quais sejam: cartas de clientes, e-mail, transcrição de gravações ou de chamadas telefônicas, contratos, documentação técnica, patentes, notícias, artigos e páginas de Internet (TAN, 1999).

No caso do sistema de SAC, a informação registrada é textual, seja na transcrição da mensagem enviada por telefone, quanto no preenchimento dos formulários específicos por meio da Internet.

Documentos textuais que fazem sentido aos seres humanos não são apropriadamente organizados para a mineração de dados em computadores. Os documentos, em seu estado natural, necessitam de pré-processamento antes de sua manipulação ou mineração computadorizadas para a descoberta de padrões e relacionamentos entre os documentos da coleção. Embora a mente humana reconheça capítulos, parágrafos e sentenças, os computadores requerem dados que estejam organizados na forma de matrizes com linhas, colunas e contagem de freqüências.

Dada a estrutura de dados disponíveis na forma de textos, a pesquisa em descoberta de conhecimento em textos é uma necessidade para que a ciência e, conseqüentemente, as organizações possam tirar proveito da enorme quantidade de informação potencialmente útil e desconhecida. Dessa forma, as instituições poderão se valer do patrimônio textual que não é utilizado por falta de ferramental apropriado e transformá-lo em vantagem competitiva.

Apesar de mais de uma década de pesquisa nesta área, o assunto em língua portuguesa permanece restrito a alguns grupos de pesquisa nas universidades e pouco foi escrito no nosso idioma. Na literatura em inglês já se pode buscar referências para que as teorias e técnicas sejam adaptadas ao português, especialmente, do Brasil.

A base de dados do SAC é rica em informações sobre a satisfação do cliente em relação à organização, produtos e serviços, porém permanece com o acesso dificultado pela forma textual em que são registrados os atendimentos e, neste sentido, de difícil utilização para a gestão estratégica da empresa.

A aplicação da DCT possibilitará a elaboração de indicador que complementará a visão dos tomadores de decisão em relação aos consumidores, com o objetivo de tornar-se mais uma ferramenta de gestão que seria transformada em vantagem competitiva.

Do ponto de vista da operação, a automatização de procedimentos operacionais, que atualmente são realizados por pessoas, contribuirá na liberação destes profissionais para atividades intelectuais ou cognitivas, exclusivas do ser humano, e para a definição da política de atendimento mais eficiente.

5 REVISÃO DE LITERATURA

5.1 Serviço de Atendimento ao Consumidor

A satisfação do cliente nem sempre foi o foco das empresas que ofereciam seus produtos conforme as suas necessidades e padrões de negócios. Kotler (1972) introduziu a satisfação do cliente como elemento fundamental, na teoria do Marketing, para a sustentabilidade das organizações.

No século XX, principalmente após a década de 70 e nos países desenvolvidos, o consumidor foi amadurecendo e provocou movimentos sociais que buscariam a ampliação de seus direitos através de demandas judiciais obrigando as empresas a reverem suas posturas.

No Brasil, se comparado aos países mais desenvolvidos, vive-se um período relativamente curto de proteção garantida pelo Estado aos clientes das empresas que se estabeleceram por todo país.

Apenas em 1991, o Código de Defesa do Consumidor foi promulgado para atender ao desejo da sociedade de garantir o seu direito no ato de consumir e poder contestar aquilo que lhe foi vendido, caso não satisfizesse as suas expectativas.

Nesse contexto, houve uma reorientação com foco no cliente e na sua satisfação como garantia de fidelização da clientela, sustentando, assim, a lucratividade em longo prazo (KOTLER, 1972).

No início, a insatisfação de consumidores, cunhada pelo termo “consumerismo” que vem da terminologia americana *consumerism* (BREDER, 2002), foi a origem do movimento social que desencadeou inúmeros estudos no sentido de compreender o cliente para oferecer-lhe mais e melhor.

Segundo Zülzke (1997), o consumerismo apresentou-se como um movimento que questionava a produção em massa, as técnicas de marketing e as relações de confiança entre vendedores e compradores.

Chauvel (2000) afirma que a reclamação foi analisada por várias disciplinas e o conceito mais aceito atualmente é que a satisfação é uma avaliação, efetuada *a posteriori*, relacionada à determinada transação.

Diante disso, a reclamação pode ser definida como a expressão de frustração da expectativa de consumidores e, neste sentido, é uma grande oportunidade, oferecida à empresa, para realinhar seus produtos e serviços com foco no cliente e reforçar o vínculo cliente-empresa (BARLOW; MOLLER, 1996).

Com a força dos movimentos sociais, a imposição judicial e a percepção de algumas empresas de que a reclamação poderia ser um excelente negócio, a mudança de foco do mercado foi inevitável e a criação de canais, pelas empresas, para “ouvir” o cliente foi uma saída de grande potencial econômico.

O consenso entre as empresas para incentivar o cliente a manifestar sua visão em relação às organizações levou à criação dos Serviços de Atendimento ao Consumidor (CHAUVEL, 2000) como canal de comunicação que ajudaria as empresas a corrigirem produtos, serviços e a própria estratégia dentro do mercado. Ainda, segundo a autora, o simples fato de existência do SAC requer mudança de postura, isto é, mais abertura e predisposição ao diálogo.

Entretanto, deve-se esclarecer que a criação de SAC cosmético[1], sem compromisso com os anseios do consumidor, em longo prazo, pode ser dramática em termos de resultados, pois a clientela se sentirá subestimada e sua reconquista se tornará muito mais difícil, mesmo que seu produto seja equiparado à concorrência (ZÜLZKE, 1997).

Nesse contexto, a empresa muda internamente para visualizar o ponto de vista do consumidor e assim estabelecer um novo padrão de atendimento, pois segundo Barlow e Moller (1996) as reclamações são um presente ao planejamento estratégico das empresas. Ainda, elas se constituem de informações de baixo custo e sem intermediários no processo de comunicação cliente-empresa.

Assim, planejar com base na reclamação significa relacionar-se melhor e tratar clientes distintos de forma distinta, pois ignorar suas diferenças, não as

[1] Somente de aparência,isto é, não há o desejo de realmente estabelecer o canal de comunicação no sentido do cliente para empresa.

eliminam, nem tornam os clientes iguais, pois eles serão sempre diferentes (PEPPERS; ROGERS, 2000).

A relação entre a reclamação e o lucro tem sido objeto de estudo e as pesquisas informam que 90% de clientes insatisfeitos não reclamam, mudam de empresa (BREDER, 2002).

A participação dos SAC nos negócios parece ser um caminho viável para o estabelecimento de relação de confiança como nos ensina Sheltman (1999) que apenas a coleta de informações, a troca de produtos, o ressarcimento do cliente e a perda do cliente não é mais tolerado no mercado, pois o cliente requer um tratamento diferenciado e personalizado. Conseqüentemente, o que diferencia uma empresa de outra é a maneira como cada uma realiza o atendimento e o estabelecimento de um ciclo de ações para satisfazê-lo de forma mais personalizada.

Por fim, no mundo dos negócios onde a concorrência é inclemente, a variedade de opções abre um leque enorme de opções ao consumidor e já não provoca tanto efeito. Em tempos de massificação, os consumidores querem que seus nomes sejam conhecidos e que suas diferenças sejam observadas pelas empresas, isto é, que o atendimento seja sob medida. Uma porta aberta para manifestação do consumidor é o SAC, porém há que se extrair o máximo que ele pode dar: informação na medida do cliente.

5.2 Descoberta de Conhecimento em Dados

A análise da literatura sobre DCT revela que essa área tem sua origem na aplicação de técnicas de Descoberta de Conhecimento em Dados na informação textual. Diante dessa constatação, é importante, inicialmente, conceituar e contextualizar a DCD.

Com o advento da digitalização de documentos e o desenvolvimento das redes, o volume de informação aumenta além da capacidade humana de apreensão e, dessa forma, existe um lapso crescente entre a criação de dados e a compreensão deles (FRAWLEY ET AL, 1992).

Existe uma necessidade premente de soluções que permitam a extração de informação útil desse universo digital em rápida expansão, pois segundo Berry e Linoff (1997), vive-se o paradigma de gigabytes (GB) de dados, porém nenhuma informação.

À medida que a tecnologia evolui, esforços são dedicados para viabilizar a utilização mais eficiente de grandes volumes de dados e da aquisição da informação ainda por ser decifrada. A DCD[2] apresenta-se como uma opção para atender a essa necessidade.

A DCD vem sendo vendida pela indústria de software como mineração de dados, ou em inglês, data mining, porém, neste trabalho, adotar-se-á o termo DCD por ser mais abrangente. O processo de descoberta de conhecimento em dados compreende a seleção de dados, o pré-processamento que envolve sua adequação aos algoritmos, a efetiva mineração de dados, isto é, o uso de técnicas de mineração, a validação dos resultados e, finalmente, a análise e interpretação dos resultados para a aquisição do conhecimento. As etapas desse processo serão discutidas adiante.

Para Fayyad et al (1997), em geral, o campo de pesquisa da DCD preocupa-se com o desenvolvimento de métodos e técnicas que buscam trazer sentido aos dados. Seu processo básico é traduzir a informação do seu nível mais elementar, o dado, geralmente armazenado em grandes volumes, em formas mais compactas, mais resumidas e mais úteis. Os métodos tradicionais de transformação de dados em informação situam-se na análise manual e na interpretação, porém, em contraste com a farta disponibilidade de bases de dados, tornam-se lentos, caros e altamente subjetivos. Assim, a DCD é uma tentativa de lidar com um problema que, na era da informação digital, tornou-se real para todos nós: a sobrecarga de informação.

Frawley et al (1992) afirma que a Descoberta de Conhecimento é a extração não-trivial da informação implícita, nos dados, previamente desconhecida e potencialmente útil e Fayyad et al (1997) complementa que a DCD é o processo de descoberta não-trivial, em dados, de padrões válidos, novos, potencialmente úteis e,

[2] Do inglês: Knowledge Discovery in Database (KDD), existe uma discussão entre diversos autores da área a respeito da abrangência do termo KDD e Data Mining (DM), porém os termos são referidos em vários trabalhos indistintamente.

finalmente, compreensíveis. Dessas afirmações, entende-se que dado é um conjunto de fatos e padrão é a estrutura implícita que será encontrada. O termo processo envolve a preparação dos dados, a busca por padrões, a avaliação do conhecimento descoberto e os refinamentos necessários em repetidas iterações. Pelo termo não-trivial depreende-se que a busca ou inferência não seja uma operação direta de quantidades pré-definidas, como por exemplo, o cálculo de uma média. Além disso, que os padrões descobertos sejam válidos em novos dados com algum grau de confiabilidade. Deseja-se, ainda, que a descoberta seja uma novidade que agregue alguma utilidade e benefício ao usuário e, por último, que seja compreensível, mesmo que necessite de pós-processamento. (FAYYAD ET AL, 1997).

Segundo Berry e Linoff (1997), a DCD é a análise e exploração automáticas ou semi-automáticas de grandes quantidades de dados com o objetivo de descobrir regras e padrões significativos.

As definições acima exprimem visões com nuances sobre o mesmo tema, pois enquanto a segunda privilegia os aspectos computacionais, quais sejam, os algoritmos dedicados à mineração de dados, bem como o poder da máquina na execução de tarefas de manipulação de grandes volumes de dados e sua transformação em informação capaz de ser utilizada pelo homem, a primeira trata do processo de descoberta como um todo, isto é, desde a aquisição do dado, seu armazenamento, a mineração que retira a informação codificada até a sua apresentação ao usuário final. Visto dessa forma, a técnica pode ser assemelhada à automatização do ciclo informacional da Ciência da Informação.

Uma outra definição do termo foi apresentada por Han e Kamber (2000) que ensinam que a DCD refere-se à extração ou mineração de conhecimento em grandes bases de dados, ou ainda, o processo de encontrar pequenos grupos de pepitas[3] preciosas em grandes quantidades de dados.

Para Hand et al (2001), a DCD é a análise de bases de dados, freqüentemente grandes, com o objetivo de achar relações insuspeitas e resumir os dados em novas maneiras que sejam compreensíveis e úteis ao usuário. Destaca-

[3] Os autores utilizam essa metáfora para traçar um paralelo entre a extração da informação e a do ouro que, em seu estado natural, freqüentemente está encrustado na pedra ou no cascalho e necessita de mineração para separá-lo.

se, ainda, que as bases de dados utilizadas pela DCD foram em muitos casos coletadas para outros propósitos, como, por exemplo, uma base de transações realizadas nos caixas de uma rede de supermercados que tem por objetivo inicial o controle contábil, mas que pode ser utilizada para descoberta de padrões de consumo. Nesse sentido, as relações encontradas pelo processo da DCD seriam novas, como se deseja, e trariam diferentes percepções ao usuário.

Agrawal e Psaila (1995) afirmam que a DCD é a descoberta eficiente de padrões previamente desconhecidos em grandes bases de dados. Essa afirmação não significa que será verificada simplesmente a existência de padrões nos dados e, sim, que serão realmente descobertos.

Portanto, o objetivo da DCD é encontrar padrões interessantes ocultos em grandes quantidades de dados e fornecer informações como insumo para aquisição do conhecimento. Além disso, oferece fundamentalmente novas capacidades, isto é, a habilidade para otimizar a tomada de decisão utilizando métodos automáticos para aprender com ações passadas (BERRY, LINOFF, 1997).

O processo da DCD apresenta-se como uma atividade multidisciplinar que se apropria de técnicas que vão além do escopo de uma área em especial (FAYYAD ET AL, 1997). Do ponto de vista histórico, de acordo com Kodratoff (1999), a DCD integra várias abordagens de aquisição de conhecimento de diversos campos da ciência como o aprendizado de máquina, que inclui os tipos de aprendizado simbólico, estatístico, neural, bayesiano, a tecnologia de bancos de dados, a visualização de dados, a recuperação de informação, a interação homem-máquina e a Ciência Cognitiva (FRAWLEY ET AL, 1992) (HAND ET AL, 2001).

Por fim, o termo DCD, em inglês KDD, foi cunhado por Frawley et al (1991) e a palavra conhecimento, nesse contexto, não é definida por sua visão na filosofia[4] e, sim, por considerar que os padrões descobertos possam adicionar alguma informação nova ao usuário.

A DCD vem sendo utilizada há mais de uma década e estabeleceu-se como solução que auxilia as organizações e pesquisadores em geral na transformação de

[4] Para o leitor interessado na visão filosófica, sugere-se a leitura de Hessen, J. Teoria do conhecimento: tradução João Vergílio Gallerani Cuter, 2ª ed., São Paulo, Martins Fontes, 2003.

dados em informação da qual se adquire o conhecimento. Os dados são armazenados em estruturas de banco de dados bem definidas, ou seja, os dados encontram-se em formatos apropriados para serem explorados por softwares especialistas.

Resumindo, a DCD, de acordo com a visão de Frawley et al (1992), apresenta quatro principais características:

1. Linguagem de alto-nível – necessariamente não precisa ser utilizada por humanos de maneira direta, mas que seja compreensível para eles.
2. Precisão – representação das descobertas descrevem, inclusive, as imperfeições da base de dados em estudo e, portanto, retratar o seu grau de confiança é fundamental. Essa confiança envolve fatores como integridade dos dados, tamanho da amostra utilizada, entre outros. Nesse sentido, a representação dessas imperfeições é expressa por medidas de graus de confiança objetivando ter segurança necessária para justificar o conhecimento descoberto.
3. Resultado interessante – os conhecimentos devem estar intrinsecamente orientados ao objetivo do usuário.
4. Eficiência – do ponto de vista computacional, a execução das tarefas é factível, isto é, previsível e aceitável.

O desenvolvimento da DCD está intrinsecamente relacionado à evolução da tecnologia. A DCD vem sendo consolidada como um poderoso ferramental para auxiliar o homem na exploração da grande quantidade de informação disponível em formato eletrônico, dadas as limitações humanas no manuseio e interpretação dessa informação.

5.2.1 O Processo da DCD

Observando-se pontos de vista expressos em vários trabalhos que enfocam a DCD, fica claro que ainda não existe consenso tanto na sua definição, quanto na sua constituição, isto é, o processo que a compõe. Entretanto, vários autores sugerem processos para a implementação do ciclo da DCD e, nesse sentido, não há a melhor

abordagem. Sintetizando as visões de Fayyad et al (1997), Berry e Linoff (1997), Trybula (1997), Han e Kamber (2000) e Kantardzic (2003), observam-se variações na ordem de apresentação das etapas que poderiam ser resumidas nos passos seguintes:

- Compreensão e definição do problema: estabelecimento do objetivo e levantamento de obstáculos para o desenvolvimento da atividade desejada, além das áreas ou pessoas que poderão ser beneficiadas com o valor agregado da informação extraída dos dados.
- Seleção de fontes de dados: adaptação da base de dados com a coleta, a combinação e a integração de fontes com objetivo de atender aos objetivos da DCD.
- Processo de limpeza e adequação dos dados: remoção de erros ou inadequações de campos e a transformação de dados com objetivo de ajustá-los à DCD. Esse passo é freqüentemente citado como pré-processamento.
- Análise exploratória: exame dos dados buscando estruturas que expressem alguma relação entre variáveis ou registros.
- Redução de variáveis: redução da dimensionalidade em função da quantidade de variáveis. Em muitos casos o número de variáveis é tão grande que inviabiliza a análise.
- Relacionamento de objetivos: a escolha do algoritmo para mineração de dados apropriado à intenção do usuário ou do analista.
- Mineração de dados: uso específico da ferramenta de mineração que tem o propósito de extrair padrões nos dados.
- Interpretação de resultados: a análise dos padrões para verificação de sua utilidade e realimentação de informação.
- Transformação do conhecimento adquirido em ação: utilização do conhecimento extraído ou sua incorporação à base de conhecimento acumulado.

A figura 1 apresenta um esquema gráfico que contempla esses passos. Observa-se que na fase de avaliação ou interpretação, o processo pode ser retomado em qualquer um dos passos anteriores, de acordo com o julgamento do especialista.

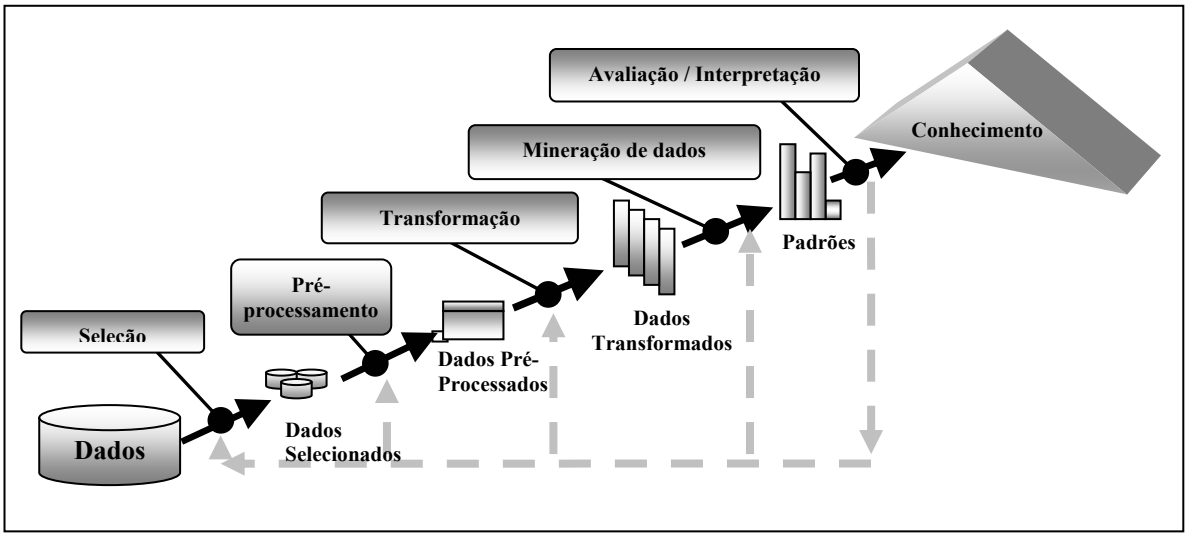

Figura 1 – Passo a passo que compõe o processo de DCD.

Fonte: Fayyad et al, 1997, pág. 41

A DCD é uma combinação interdisciplinar de técnicas e métodos que foi concebida para resolver problemas, na sua grande maioria, em dados numéricos e estruturados, isto é, armazenados em bancos de dados. Contudo, apenas uma pequena parte dos dados está adequada ao tratamento proposto pela DCD e, nesse sentido, os dados expressos em linguagem natural não são contemplados. Portanto, existe uma necessidade de adaptação da DCD para que a linguagem natural seja passível de processamento automático visando à extração de conhecimento.

5.3 Processamento de Linguagem Natural – PLN

Com a evolução tecnológica, houve uma massificação de textos de toda ordem, seja em correspondências eletrônicas, em publicações científicas ou em sítios na Internet, com diversos propósitos.

Segundo Hearst (1999), o texto expressa uma fonte de informação tão vasta, quanto rica, porém codificada de maneira que é difícil de ser decifrada

automaticamente. Assim, a ciência vem buscando soluções para simular a cognição humana que processa o texto e o apreende de maneira satisfatória.

De acordo com Manning e Schütze (1999), o estudo da Lingüística vem contribuir para resolver esse problema, pois busca caracterizar e explicar a diversidade de observações lingüísticas que nos cerca, seja em diálogos, seja na escrita, seja em qualquer outro meio. Uma parte preocupa-se com o lado cognitivo de como o homem adquire, produz e entende a linguagem, outra parte, a compreensão da relação entre discurso lingüístico e o mundo e, a terceira, com a compreensão das estruturas lingüísticas pelas quais o homem se comunica.

Paralelamente, o desenvolvimento da informática tem possibilitado grandes avanços no estudo das línguas naturais. A área que examina as relações entre a Lingüística e a Informática é a Lingüística Computacional que objetiva a construção de sistemas especialistas em reconhecer e produzir informação em linguagem natural. Encontram-se neste contexto os estudos de PLN que têm por objetivo a interpretação e geração de informação nos diferentes aspectos da língua: sons, palavras, sentenças e discurso nos níveis estruturais, de significado e de uso. (VIEIRA; LIMA, 2001).

O PLN já existe há décadas e, nesse ínterim, desenvolveram-se várias técnicas tipicamente lingüísticas, isto é, as sentenças do texto são separadas em partes gramaticais (sujeito, verbo, etc) utilizando uma gramática formal ou um léxico, então a informação resultante é interpretada semanticamente e usada para extrair informação sobre o que foi escrito (KAO; POTEET, 2005).

Não se propõe aqui uma discussão detalhada do PLN, seus métodos e suas técnicas e, sim, a contextualização da relação entre a DCD e a DCT. Assim, durante várias décadas, inúmeras pesquisas têm provocado avanços no PNL e, atualmente, encontram-se procedimentos disponíveis capazes de realizar o tratamento do dado textual de maneira a possibilitar sua transformação e sua estruturação na forma adequada ao uso pela DCD. Alguns desses procedimentos são ferramentas essenciais para viabilizar a Descoberta de Conhecimento em Textos – DCT.

5.4 Descoberta de Conhecimento em Textos

Como enunciado anteriormente, os métodos da DCD lidam com dados em formatos bem estruturados isto é, em formatos numéricos ou simbólicos preparados para a leitura por computador (WEISS et al., 2005). Por outro lado, Tan (1999) e Feldman et al (2001) declaram que 80% da informação de empresas está em documentos textuais e, decorre daí, que existe grande espaço para pesquisa.

De acordo com Dörre et al (1999), a informação textual não está prontamente acessível para ser usada por computadores, ou seja, ela é apropriada para que pessoas, através da leitura e dos processos cognitivos característicos dos humanos, manipulem e apreendam as informações contidas nesse formato.

Ainda que as aplicações da DCD não estejam consolidadas, localizam-se em patamar altamente especializado e, para algumas formas de análise, já se encontram em fase madura (WEISS et al, 2005). Dessa maneira, o problema a ser resolvido pela DCT seria a adequação de dados textuais, isto é, o processamento da linguagem natural, para utilização das técnicas da DCD.

Para Trybula (1999), a DCT assemelha-se a DCD com exceção ao foco em coleções de documento textuais. Essa visão está de acordo com Feldman et al (2001) que afirma que a DCT é a área dentro da DCD que se concentra na descoberta de conhecimento em fontes de dados textuais.

Segundo Tan (1999), a DCT pode ser vista como uma extensão da DCD, pois se refere ao processo de extração de padrões não-triviais e de conhecimento útil para determinado objetivo em documentos não-estruturados. Todavia, a tarefa da DCT torna-se mais complexa em função da manipulação de dados textuais registrados em linguagem natural.

Na realidade, vários autores afirmam que as bases textuais apresentam-se de forma não-estruturada. Porém, possuem uma estrutura implícita que necessita de técnicas especializadas para ser reconhecida por sistemas automatizados. O processamento de linguagem natural (PLN) trata exatamente da descoberta destas estruturas implícitas, como por exemplo, a estrutura sintática (RAJMAN; BESANÇON, 1997).

A integração de técnicas de PLN e DCD constitui a Descoberta de Conhecimento em Texto que objetiva automatizar o processo de transformação de dados textuais em informação para possibilitar a aquisição do conhecimento.

Loh et al (2000) entende a DCT como a aplicação de técnicas e ferramentas com o propósito de buscar conhecimento novo e útil de coleções de textos. Esse é também o entendimento de Kodratoff (1999) que acrescenta que o conhecimento extraído é o conjunto de padrões, compreensíveis aos humanos, resultantes do processamento da informação textual.

De acordo com Wives e Loh (1999), a DCT é a evolução da Recuperação de Informação (RI), visão compartilhada por Dörre et al (1999), pois enquanto na RI o usuário sabe previamente o que está buscando, na DCT as informações serão extraídas em um conjunto de textos que deverão apresentar ao usuário um conhecimento útil e novo, isto é, que satisfaça sua necessidade de informação. Essa visão é reafirmada por Trybula (1999) declarando que, diferenciando-se da RI, a DCT apresenta a informação sobre tópicos relacionados, sem necessariamente responder a uma pergunta específica previamente definida pelo usuário.

A figura 2 apresenta o ciclo do processo da DCT e pode ser visto como uma consolidação das concepções de autores como Wives e Loh (1999), Dörre et al (1999) e Tan (1999) que equivale a uma adaptação ao modelo proposto por Fayyad et Al (1997).

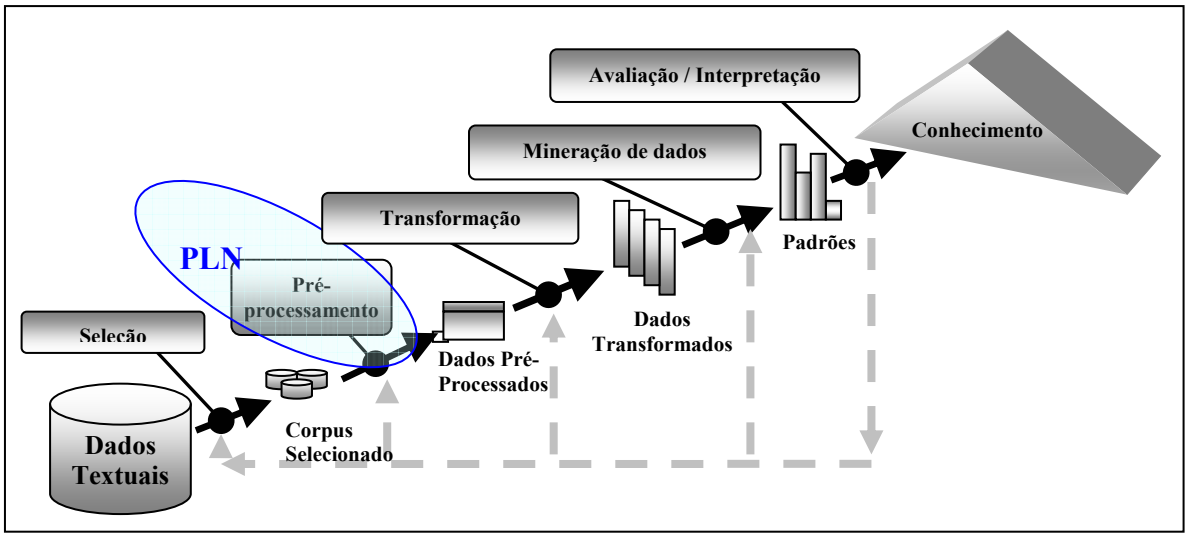

Figura 2 – Processo de DCT. Adaptação de Fayyad et al (1997), pág. 41.

De acordo com a figura 2, observa-se que o processo de DCT abrange a seleção do corpus, o pré-processamento, que envolve sua adequação aos algoritmos, a efetiva mineração de dados textuais, a validação dos resultados e, finalmente, a análise e interpretação dos resultados para a aquisição do conhecimento. Comparando-se esse esquema ao apresentado na figura 1, verifica-se que as fases da DCT seriam idênticas às da DCD exceto pela utilização das técnicas de PLN para a transformação do texto em formato adequado para a mineração dos dados.

Uma diferença entre a DCT e DCD está na transformação de dados: a DCD utiliza algoritmos que procuram associações ou correlações entre grupos de dados e objetivam descobrir novas conexões. A DCT fornece associações ou correlações similares de ocorrências de padrões de textos, contudo, o usuário deve decidir se essas descobertas têm realmente significado e essa tarefa exige uma significativa compreensão do conteúdo da informação (TRYBULA, 1999).

Vale a pena destacar que, embora a aderência entre o PLN e DCT seja muito forte, seus objetivos não se confundem, pois o PLN trata dos aspectos lingüísticos relacionados a um texto específico, enquanto que a DCT busca as relações contidas entre os textos de uma coleção com o objetivo de apresentar a informação relacionada a um grupo ou grupos de textos. Ainda, o PLN analisa o conteúdo dos textos e, por outro lado, a DCT se utiliza dessas análises, em uma fase de pré-processamento, para transformá-las em dados apropriados para a descoberta de padrões e conhecimento entre os textos da coleção (KODRATOFF, 1999).

As aplicações da DCT são variadas e possibilitam a manutenção do ciclo informacional com mais eficiência, pois automatizam tarefas que exigem muito tempo do ser humano. Segundo Weiss et al (2005), as principais aplicações utilizadas são:

1. **Classificação de Documentos**
    a. Agrupamento de textos - técnica que permite criar novas categorias de textos a partir de uma coleção;

b. Categorização de textos - técnica que permite classificar os textos de uma coleção em categorias pré-definidas. Ex. Filtro de e-mail, Inteligência Competitiva utilizando informação disponível na Web;

2. Recuperação da Informação

a. Motores de busca equipados com técnicas de DCT para melhorar a precisão. Nesse caso, a RI é considerada um caso particular da categorização de textos, na qual os textos são classificados pelo critério de similaridade a partir de um documento que está sendo considerado ou do agrupamento em que será alocado.

3. Extração da Informação

a. Extraem-se do texto palavras relevantes ao assunto pesquisado e seus termos correlacionados. Ex. Notícias sob demanda (*clipping*), temas científicos associados a autores ou instituições.

4. Predição e Avaliação

a. Tipo de técnica bem conhecida no meio financeiro, pois os mercados necessitam de métodos preditivos para fazerem suas apostas. No caso do texto, os programas de aprendizagem utilizam modelos bem definidos para “aprenderem” como é o documento padrão e elaboram regras para fazer previsões baseadas nos novos documentos. Ex. Boletins da Bolsa de Valores.

A interação homem-máquina na DCT proporciona vantagens no sentido de que as operações repetitivas podem ser executadas pela máquina com velocidade muito maior e, dessa maneira,o homem pode dedicar-se a tarefas analíticas que demandam um nível de compreensão humana potencial ainda não alcançado no universo da Inteligência Artificial (IA).

Dessa maneira, a DCT possibilita então o reconhecimento e a produção da informação apresentada em linguagem natural e, nesse sentido, vem contribuir enormemente com a Ciência da Informação no que tange ao tratamento e recuperação da informação.

5.4.1 Fundamentos

A Ciência da Informação vale-se de suas teorias, metodologias e tecnologias de análise e manipulação estrutural para tornar a informação registrada disponível e, para tanto, volta-se para a compreensão de sua natureza e uso social através de métodos qualitativos e quantitativos.

Todo documento na acepção de informação registrada está sujeito a diferentes abordagens, entretanto seria razoável apontar duas direções complementares e interdependentes: uma direcionada para o conteúdo enquanto tal e a outra para a estrutura do próprio documento (MIRANDA, 2003).

Nesse sentido, a DCT pode ter uma abordagem voltada tanto para a compreensão do conteúdo, quanto para a análise da estrutura de documentos, isto é, análise estatística descritiva, e ambas visam identificar padrões implícitos em uma grande coleção de documentos.

Segundo Woodfield (2004), é importante explorar essas duas vertentes de análise e, para tanto, alguns conceitos essenciais são necessários para o aprofundamento do tema:

- *Token* é uma sucessão contígua de caracteres que não contém um separador. Um separador é um caractere especial tal como um espaço em branco ou sinal de pontuação.
- Termo é um *token*, ou mais, com significado específico numa dada linguagem.
- Documento consiste em um grupo de *token*.
- Corpus é uma coleção de documentos.

Uma abordagem para descrição de um corpus pode ser o reconhecimento das características físicas dos documentos. Tais características podem incluir o tamanho do documento em relação ao número de palavras, sentenças, caracteres, distribuição do comprimento de palavras, a contagem de freqüência absoluta e relativa de palavras-chave e assim por diante. Esse tipo de análise contribui para

identificar características que separam ou distinguem um documento de outros em uma coleção.

5.4.2 Características de um Documento

Um documento consiste essencialmente de elementos como letras, palavras, sentenças, parágrafos, pontuação e possíveis itens estruturais tais como capítulos e secções.

Esses elementos podem ser contados como, por exemplo, o número de caracteres, palavras e sentenças ou, ainda, resumidos na forma de medidas estatísticas como média, mediana ou variância.

Medidas estatísticas de características físicas podem ser úteis na identificação de autoria, entretanto, limitar a análise somente a esses valores, significa ignorar as características lingüísticas de documentos. Tais características são dominantes na identificação de documentos e, em conseqüência disso, somente resumos estatísticos são raramente empregados na DCT.

Resumos estatísticos, que ignoram características lingüísticas, convertem textos não estruturados em vetores numéricos estruturados. Entretanto, satisfazer a meta de converter texto em números não resolve o problema empregando-se apenas resumos estatísticos. Há que se considerar a quantidade de informação perdida quando se ignoram as palavras e seus significados em determinados contextos.

Esses procedimentos de conversão do texto em números são poderosas ferramentas para exploração dos dados que podem fornecer informações importantes para adequação do corpus e da escolha de técnicas para o processo de descoberta. Aliados a isto, muitos esforços são direcionados à extração de conceitos e significados do texto nos quais a DCT resolve, em parte, os problemas de somente considerar as características quantitativas do texto.

No processo de conversão do texto em elemento adequado aos algoritmos computacionais é necessário que o texto seja, inicialmente, transformado em unidades de informação que possam, ao mesmo tempo, conter a informação original do texto e estar apropriada ao manuseio por técnicas estatísticas ou computacionais.

Essa extração de unidades de informação foi apresentada por Wakefield (2004) e complementada por Woodfield (2004) segundo a ordem de complexidade, como se segue:

Tabela 1 – Extração de unidades de informação por ordem de complexidade

Atividade	Descrição	Exemplo
Extração de *Token*	o texto é separado em palavras sem considerar seu significado	Ciência da Informação ⇒ciência, da, informação
Extração de Termos	o texto é separado em palavras considerando seu significado dentro de uma linguagem específica.	*Token* + linguagem específica ⇒ termo (mineração de texto)
Extração de Conceitos	conceito indica assuntos ou tópicos contidos em um texto e a extração de conceitos revela somente se tais conceitos estão presentes em um texto sem que haja a preocupação com seus detalhes	Um artigo pode tratar de astronomia, carros ou câncer
Extração de Entidades	As entidades representam pessoas, localidades ou coisas dentro de um texto, freqüentemente, na forma de substantivos	Pessoa ⇒ Senhor Coelho Animal ⇒ coelho
Extração de fato pontual	um fato pontual relaciona uma entidade a uma ação. Esses pares podem estar separados por outros termos e, portanto, devem extraídos por sistemas capazes de reconhecer os fenômenos lingüísticos	sujeito ⇒ ação terrorista ⇒ explodiu
Extração de fato complexo	um fato complexo é relação ou relações de múltiplas entidades e qualificadores	compreensão de linguagem natural

No contexto da DCT, a extração de unidades de informação representa um ou mais passos no processo de descoberta e estão destacadamente relacionados a uma linguagem específica. Importante observar também que a PLN empresta alguns algoritmos para solucionar problemas relacionados à ambigüidade presente na grande maioria dos textos.

5.4.3 Processamento Textual

Muitos desafios estão intrínsecos ao dado textual, uma simples mensagem enviada por correio eletrônico compõe-se de variáveis difíceis de serem controladas sob o ponto de vista de máquinas. Por exemplo, um texto mal formulado com idéia confusa, erros ortográficos, abreviaturas fora do padrão, jargão técnico e peculiaridades do autor contribuem para aumentar a complexidade na tarefa de decodificação automática (WOODFIELD, 2004). O sucesso na remoção de ambigüidades na fase de preparação dos dados está diretamente relacionado ao êxito no tratamento dos algoritmos de mineração de texto.

Com o objetivo de adequar o processo de DCT ao de DCD, a preparação de dados é tarefa essencial e crítica que demanda grande parte do tempo gasto em todo processo de descoberta. Essa adequação permite que o texto possa ser representado na forma adequada aos vários algoritmos da mineração de dados.

Todo o processo de preparação dos dados textuais necessita de um ou mais tratamentos que se apresentam a seguir.

5.4.3.1 Coleta de Dados

A coleta de dados é, sem dúvida, a primeira tarefa a ser realizada, porém os acervos de documentos relevantes ao estudo podem ser conhecidos ou se constituírem como parte do problema.

Segundo Weiss (2005), se os documentos já estão identificados, então eles podem ser obtidos e a principal tarefa é efetuar a eliminação de ruídos e assegurar que a amostra é de boa qualidade. Assim como nos dados não textuais, a intervenção humana pode comprometer a integridade dos dados no processo de coleta, por isso requer-se extremo cuidado nesta tarefa.

A leitura automatizada de fontes textuais implica em dificuldades práticas em relação aos textos que podem se apresentar em vários formatos e idiomas. Quando estendemos essa tarefa, que parece simples, ao domínio da Internet, a complexidade pode ser multiplicada várias vezes.

5.4.3.2 Padronização de Documentos

Após o processo de coleta, a definição do formato que será utilizado para todos os documentos é tarefa importante para a estratégia de descoberta, visto que alguns aplicativos necessitam de formatos pré-definidos.

A dependência da língua em que foi escrito o documento é outro fator bastante relevante, uma vez que se tratam de procedimentos que utilizam a lingüística como base. Portanto, agrupam-se os textos do mesmo idioma para serem estudados separadamente.

5.4.3.3 Transformação do Texto em *Token* (*tokenization*)

Um texto possui um fluxo ordenado de palavras que seguem as normas lingüísticas de um idioma para que ele faça sentido para o leitor. No entanto, para o propósito de manipular o texto com computadores no sentido de extrair suas características, o processo utilizado é de separação do texto em unidades chamadas tokens.

Segundo Manning e Schultz (1999), esses tokens apresentam-se como palavras, números ou sinais de pontuação extraídos do texto. Importante observar que os sinais de pontuação podem trazer informação sobre a macro estrutura do texto e por isso não devem ser negligenciados.

Geralmente, o que diferencia um token do outro são os espaços entre eles e freqüentemente os algoritmos que executam a divisão do texto em tokens utilizam o espaço como delimitador. Aqui, também, se requer cuidado na execução da tarefa, pois temos, em português, palavras compostas que quando separadas possuem significados diferentes, por exemplo, “Casas Bahia” representa uma conhecida rede de lojas, contudo, se lida isoladamente, a palavra “casas” ou “Bahia” não possuem relação alguma com comércio de eletrodomésticos ou móveis.

Para obtenção de melhores resultados, deve-se adequar o programa que executa o trabalho de separação dos termos em função do texto que será tratado, caso contrário, muito trabalho deverá ser executado nos tokens adquiridos.

5.4.3.4 Normalização de Conteúdo

Além de formatos e idiomas, os textos devem passar por uma análise no sentido de normalizar os seus conteúdos e de padronizar toda coleção de dados para a mineração, alguns destes procedimentos são:

- Conversão de abreviações não padronizadas em termos válidos para evitar a ambigüidade. Ex. Gal. ⇒ Galicismo, Gal. ⇒ Galego;
- Inferência de potenciais problemas em relação a mesma palavra escrita, ora em letras maiúsculas, ora em minúsculas. Isto é necessário, pois a representação interna dos computadores, isto é, a representação em bytes é diferente, como por exemplo, Documento ≠ documento.
- Identificação de seções para documentos bem estruturados. Ex. Introdução, referência, sumário;
- Identificação de palavras-chave em documentos
- Identificação e rotulação de expressões sinônimas: "carro", "caminhão", "ônibus", "veículo".
- Tratamento de sinais de pontuação que podem indicar a macro estrutura textual e, portanto, não devem ser simplesmente descartados. Ex. “**,**”, ”**.**”, ”**;**”, “**:**”, “**<**”, “**>**”, “**!**”, “**?**”.
- Compilação de estatísticas dos termos para fins de exploração. Ex. comprimento de termos, freqüência de termos, freqüência de combinação de termos.

5.4.3.5 Criação de Dicionários de Apoio e Tesauros

Objetivando o tratamento adequado de alguns problemas intrínsecos da língua são criados dicionários de apoio, tesauros, listas de termos não relevantes ou específicos do jargão técnico para dar suporte ao trabalho de processamento do dado textual. Esse passo é dependente do objetivo do projeto que pode fazer uso de todos os elementos de apoio ou somente de alguns de acordo com a necessidade.

No momento em que se transforma o texto em termos individuais ou compostos observa-se que alguns aparecem muitas vezes, outros medianamente e outros raramente.

A utilização de alguns termos em detrimento de outros é uma escolha feita pelos analistas que conduzem o processo de descoberta e, para tanto, são criados dicionários especializados e listas de termos que apóiam o trabalho de escolha dos termos que serão utilizados pelos algoritmos de mineração de texto.

Uma preocupação natural seria unificar todas as palavras que possuem o mesmo significado. Então, cria-se um dicionário de sinônimos ou um tesauros que converte os termos sinônimos em um termo preferido. Esse tipo de tratamento é importante para redução da quantidade termos nos documentos, isto é, para n termos sinônimos, a máquina os trata como termos não correlacionados e, para fins de estatísticas, são computadas individualmente para cada termo, o que não é interessante para a captura do conceito do documento. Convertido os n termos sinônimos para o preferido, a análise será feita em apenas um termo que pode ampliar a sua relevância no documento.

Outra utilização destes dicionários de apoio seria correção de erros ortográficos corriqueiros. O procedimento é análogo ao dicionário de sinônimos que, neste caso, cadastram-se os prováveis erros ortográficos mais comuns. Ex. iorgute $\Rightarrow$ iogurte.

A correção ortográfica pode uniformizar o conteúdo dos textos no corpus que é importante para análise do conceito geral de documentos. Entretanto, dependendo do objetivo da análise, esse passo deve ser executado com parcimônia, pois palavras com erros ortográficos podem indicar autoria.

A datação do texto também deve ser verificada para que não seja modificado o seu conteúdo sem as devidas considerações. A escrita de determinada época não deveria sofrer correções automáticas, pois seriam descaracterizadas. Ex. Pharmácia $\Rightarrow$ farmácia .

De maneira similar, pode-se identificar abreviaturas que estão fora do padrão para serem corrigidas. Além disso, as palavras compostas não devem ser separadas ou convertidas para um sinônimo de um de seus termos, pois podem mudar de

sentido. Por exemplo, foi criado um dicionário de sinônimos que contém a relação dos termos Casa e Morada, de forma que a ocorrência do termo Morada seria convertida para o termo Casa. Não seria apropriado converter o primeiro termo da palavra composta “Casa Civil”, pois descaracterizaria o sentido. Uma solução para reconhecer esses termos automaticamente seria o cadastramento das formas compostas para sua correta identificação.

Em determinados projetos pode ser útil a identificação de Entidades que relacionam termos com categorias. O objetivo dessa tarefa é distinguir termos que contém informações bastante relevantes para o processo de descoberta.

Imaginemos um corpus contendo informações cadastrais de usuários de determinado serviço ou clientes de uma empresa. Seria de grande valia identificar no texto o nome, o endereço, o telefone, a empresa na qual trabalha e assim por diante.

Dessa forma, cria-se um dicionário de apoio contendo prováveis formatos em que são apresentadas essas Entidades. Freqüentemente, essa tarefa é realizada por analistas de domínio que detém conhecimento *a priori* das categorias que deverão ser identificadas.

5.4.3.6 Rotulação de Partes do Discurso

A língua portuguesa possui uma norma bem regulamentada e específica, as palavras são organizadas em classes gramaticais ou, mais comumente chamadas na lingüística computacional, em partes do discurso.

A tarefa é identificar e rotular se um termo é um substantivo, verbo, advérbio, adjetivo e assim por diante. Essa rotulação pode ser de grande valia na remoção de ambigüidades de termos homógrafos usados em contextos diferentes.

As sentenças:

Saquei dinheiro do **banco**. (banco significa uma instituição financeira)

O **banco** do jardim é de madeira. (banco significa objeto, um móvel)

nos mostra que a palavra “banco”, apesar da escrita idêntica, possui significados e funções gramaticais diferentes.

Dessa maneira, o mesmo termo utilizado com funções gramaticais diferentes, deve ser tratado como termo diferente tanto quanto forem as suas classificações. Na escolha do termo preferido no dicionário de sinônimos, a informação da função gramatical do termo é determinante para sua conversão, pois um termo só será convertido para o seu sinônimo se possuir a mesma função gramatical, caso contrário, deseja-se que sejam tratados como termos distintos.

5.4.3.7 Criação de Listas de Apoio

Lista de Stopwords – são listas de apoio que contém as *stopwords* que são as palavras que ocorrem com muita freqüência em uma dada língua. Um exemplo destas palavras são os artigos ou as preposições que, na maioria dos textos, possuem maior ocorrência. Uma vez que esses termos ocorrem em todos os documentos, seu poder de discriminação é muito pequeno e, por isto, eles são descartados da análise sem prejuízo de perda de informação.

Lista de Startwords – ao contrário das listas de *stopwords*, essa lista contém os termos que caracterizam o domínio do assunto a ser pesquisado. Somente os termos contidos nesta lista serão considerados. Esse tipo de lista é utilizado quando se examinam coleções altamente especializadas dominadas pelo jargão técnico e são construídas por especialistas de domínio com o objetivo de otimizar a mineração de texto.

5.4.3.8 Lematização[5]

Determinada palavra possui variações que estão estabelecidas na norma. A lematização converte todas essas variações para uma forma padrão que pode ser o radical da palavra. Segundo Baeza-Yates e Ribeiro-Neto (1999), a lematização retira do termo o afixo (prefixo e sufixo), além de reduzir suas variantes, ou seja, desinências[6] e vogal temática, ao mesmo radical que expressa um conceito comum.

[5] Freqüentemente escrito em Inglês – *Stemming*. Ressalta-se que o termo Lematização foi encontrado somente em dicionários de português europeu, como por exemplo: http://www.priberam.pt/dlpo/definir_resultados.aspx.

[6] São elementos mórficos que se apõem ao radical para assinalar as flexões da palavra (gênero, número, modo, tempo, pessoa) (TERRA, 2002)

Nesse contexto, deseja-se que o termo “livro” não seja diferente de “livros”, já que se referem ao mesmo objeto. As formas que se diferenciam segundo o gênero também podem ser reduzidas a uma única forma. Ex. Bibliotecária, bibliotecário. Ainda, as formas verbais também seriam uniformizadas segundo esse critério. Ex. é, sou, era, fui ⇒ ser

Apresentando um exemplo, teríamos os termos informação, informações, informar, informado e informando em uma única forma padrão que seria “inform”. Tal procedimento pode ser muito útil em algumas aplicações que levam em conta a quantidade de vezes que o termo se repete.

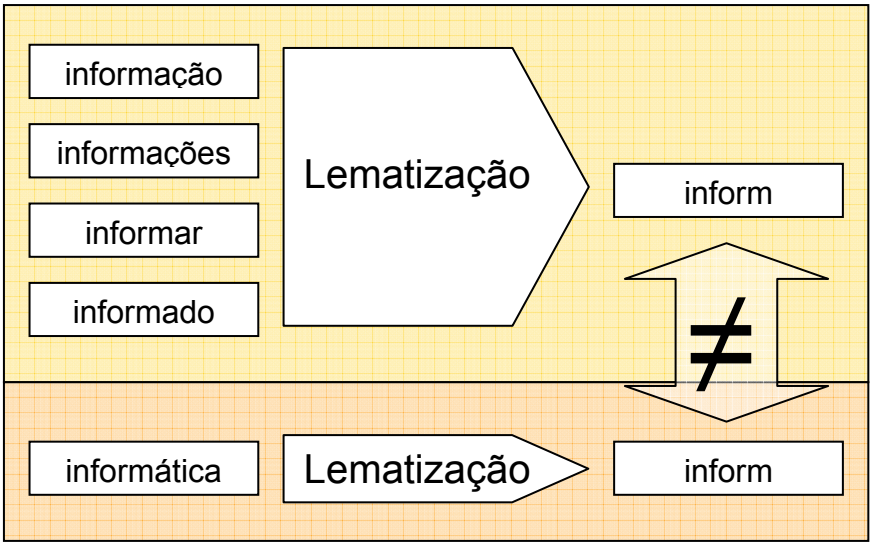

Figura 3 – Lematização de Termos

Conforme a figura 3, há que se observar que esse procedimento pode cometer alguns erros como a conversão do termo informática – que se relaciona com computação, tecnologia – para “**inform**” que se originou do termo informação. Outra consideração são os termos que não possuem formas no singular, por exemplo, costas que se for retirado o “S”, muda de sentido completamente.

Os procedimentos descritos anteriormente visam minimizar a quantidade de termos que serão criados de forma a facilitar a análise e reduzir o custo computacional por restringir a quantidade de termos que serão processados. De tal modo, essas etapas de processamento do dado textual permitirão que os documentos sejam quantificados utilizando-se as distribuições de freqüências dos termos dentro dos documentos do corpus.

5.4.4 A Representação Quantitativa do Texto

Uma parte fundamental da DCT está na adequação do texto ao formato reconhecido por algoritmos específicos para mineração, isto é, na transformação do texto em números sem que se perca a informação codificada nele.

Nas bases de dados estruturadas para a mineração, utiliza-se freqüentemente o formato de planilhas para apresentação de dados. Assim, cada registro, ou linha, refere-se a um caso e cada coluna, ou campo, refere-se a um atributo específico.

Tabela 2 – Representação de Dados

PESSOA	***IDADE***	***PESO***	***ALTURA***
Mauro	26	68	178
Iolanda	22	57	162
...	...	...	...
Sandra	40	65	153

Observa-se na tabela 2 que as linhas se referem aos indivíduos e as colunas, os atributos relacionados aos indivíduos. Todos os campos identificam o registro especifico completamente, ou seja, um indivíduo é composto pela totalidade de seus atributos. Essa é a maneira convencional de se apresentar os dados e utilizá-los para procedimentos de descoberta de conhecimento em dados.

Em um documento, os atributos podem ser representados por seus termos de maneira que possamos representá-los em tabelas nos mesmos moldes. Cada linha representa um documento de uma dada coleção e cada coluna refere-se ao termo contido no texto correspondente.

Exemplificando, tem-se uma coleção de documentos e cada documento com seu conteúdo textual, como se segue:

Doc1 – “Política é o fim”;

Doc2 – “O caminho da política é informação”;

Doc3 – “A segurança da informação é a política”;

Logo, a representação poderia ser conforme a tabela 3, como se segue:

Tabela 3 – Representação do Corpus

Documento	Termos								
	a	**política**	**É**	**o**	**fim**	**caminho**	**da**	**informação**	**segurança**
D_1	N	S	S	S	S	N	N	N	N
D_2	N	S	S	S	N	S	S	S	N
D_3	S	S	S	N	N	N	S	S	S

Nesse exemplo, tem-se uma representação binária informando a presença “S” ou não “N”. De maneira similar à tabela 2, um documento é representado pela totalidade de todos os seus atributos relacionados.

Assim, sejam ***n*** documentos, a quantidade de documentos na coleção representados por $\boldsymbol{D = \{D_1, D_2, \ldots, D_n\}}$, e m termos, ou atributos, presentes no corpus representados por $\boldsymbol{T = \{T_1, T_2, \ldots, T_m\}}$. Cada documento ***D*** é representado por ***m*** termos existentes no documento. Cada termo pode ser uma palavra simples ou composta. Então, a_{nm} representa a influência do atributo m no documento n que pode estar representada pela indicação da presença do termo, pela freqüência[7] do termo em relação ao documento ou pela freqüência do termo em relação à coleção de documentos. Portanto, para qualquer corpus poder-se-ia representá-lo conforme tabela 4.

Tabela 4 – Representação Generalizada do Corpus

Documento	*Termo*			
	T_1	T_2	…	T_m
D_1	a_{11}	a_{12}	...	a_{1m}
D_2	a_{21}	a_{22}	...	a_{2m}
...	...	...	...	...
D_n	a_{n1}	a_{n2}	...	a_{nm}

Assim, as estratégias para quantificar os termos na tabela são variadas e, em alguns casos, podem levar em conta simplesmente a existência do termo, em outros, a freqüência do termo em relação ao documento ou, ainda, a freqüência do termo

[7] Quantidade de vezes que ocorre o termo.

em relação à coleção. Cada representação privilegia uma característica em detrimento de outra, o tipo de representação é dependente da aplicação.

Representação Binária – considera-se a existência do termo. Se o valor a_{ij} é igual a 1, então o termo t_j ocorre no documento d_i. Caso contrário, a_{ij} igual a 0, para todo $j \in \{1, \ldots, M\}$ e $i \in \{1, \ldots, N\}$. Conforme apresentado na equação 1.

Equação 1

$$a_{ij} = \begin{cases} 1, \text{ se } t_j \text{ ocorre em } D_i \\ 0, \text{ se } t_j \text{ não ocorre em } D_i \end{cases}$$

Reescrevendo a tabela 4, temos a seguinte representação:

Tabela 5 – Representação do Corpus em Código Binário

Documento	Termos								
	a	**política**	**é**	**o**	**fim**	**caminho**	**da**	**informação**	**segurança**
D_1	0	1	1	1	1	0	0	0	0
D_2	0	1	1	1	0	1	1	1	0
D_3	1	1	1	0	0	0	1	1	1

Esse tipo de representação apenas informa a existência do termo, não levando em consideração a quantidade de vezes em que ele aparece. Em algumas aplicações, essa simples representação é suficiente.

Entretanto, pode ser necessário contar a ocorrência de um termo. Nesse caso, ao invés de 1 e 0, coloca-se a freqüência observada do termo referente ao documento em que foi encontrado.

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

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

Equação 2

$$a_{ij} = tf(t_j, d_i)$$

Sua representação constitui-se em:

Tabela 6 – Representação do Corpus em Freqüência

Documento	Termos								
	a	**política**	**é**	**o**	**fim**	**caminho**	**da**	**informação**	**segurança**
D_1	0	1	1	1	1	0	0	0	0
D_2	0	1	1	1	0	1	1	1	0
D_3	2	1	1	0	0	0	1	1	1

A contagem simples, conforme equação 2, leva em consideração apenas a presença do termo no documento, que não reflete como os documentos são comparados entre si dentro do corpus.

Documentos maiores tendem a ter freqüências mais altas que documentos menores e, portanto, pode ser necessário criar uma medida que leve em conta a presença do termo em relação aos outros documentos da coleção.

Nesse sentido, adiciona-se uma ponderação que leva em conta a distribuição dos termos em todos os documentos do corpus, isto é, uma medida que vai além da simples freqüência de um termo. Assim, a medida *inverse document frequency* – idf é um fator de escala para a importância do termo em relação aos outros documentos da coleção.

Equação 3

$$idf(j) = \log\left(\frac{N}{df(j)}\right)$$

na qual *df(j)* é o número de documentos que contém o termo *j* e *N* a quantidade de documentos do corpus. Onde $j \in \{1, \ldots, M\}$, e $df(j) \in \{1, \ldots, N\}$.

Logo, se *df(j)=N*, Isto é, o termo *j* ocorre em todo documento da coleção, dessa forma tem-se $\log\left(\frac{N}{N}\right) = \log(1) = 0$, portanto, essa medida favorece termos que aparecem em poucos documentos.

Então, combinando a equação 2 e 3 temos:

Equação 4

$$a_{ij} = tf(t_j, d_i) * idf(j)$$

Portanto, quando um termo aparece em vários documentos, sua importância é reduzida, pois *idf(j)* se aproxima de 0 . Caso contrário, ela é aumentada.

Geralmente, deseja-se que os documentos da coleção sejam tratados com a mesma importância, independente de seu tamanho. Essa medida possibilita que os atributos dos termos tanto para documentos maiores quanto para menores possam ser comparados na mesma escala.

Considerando que a tabela 7 tenha 3 documentos, aplicando-se a equação 4 tem-se:

Tabela 7 – Representação do Corpus em Peso

Documento	*Variáveis Termos*								
	a	**política**	**é**	**o**	**fim**	**caminho**	**da**	**Informação**	**segurança**
D_1	0,00	0,00	0,00	0,18	0,48	0,00	0,00	0,00	0,00
D_2	0,00	0,00	0,00	0,18	0,00	0,48	0,18	0,18	0,00
D_3	0,95	0,00	0,00	0,00	0,00	0,00	0,18	0,18	0,48

Existem diversas variações de ponderações aos termos da tabela de representação que podem apresentar resultados semelhantes dependendo do objetivo do trabalho.

Resumidamente, segundo Woodfield (2004), apresenta-se na tabela 8 a notação adotada e nas tabela 9 e 10 as ponderações existentes no programa SAS Text Miner utilizado nesta pesquisa.

Tabela 8 – Notação

Fórmula	*Descrição*
a_{ij}	Freqüência com que o termo *i* aparece no documento *j*
g_i	Freqüência com que o termo *i* aparece na coleção de documentos
n	Número de documentos na coleção
d_i	Número de documentos nos quais o termo *i* aparece
$P_{ij} = \frac{a_{i,j}}{g_i}$	Proporção da freqüência com que o termo *i* aparece no documento *j* em relação à freqüência do termo *i* em toda coleção

O peso total é determinado por duas ponderações chamadas Peso da Freqüência e Peso do Termo.

Considera-se o Peso da Freqüência somente em função de sua freqüência no documento, isto é, não se leva em conta a freqüência do termo em relação ao corpus. São eles:

Tabela 9 – Peso da Freqüência

Peso da Freqüência	*Fórmula*	*Descrição*
Binário	$L_{ij} = \begin{cases} 1, \text{se termo } \mathrm{i} \text{ ocorre em documento } \mathrm{j} \\ 0, \text{caso contrário} \end{cases}$	Toda freqüência tranforma-se em 0 ou 1, indicando existência ou não. Utilizado em coleções com vocabulário pequeno.
Log	$L_{ij} = \log_2(a_{ij} + 1)$	O cálculo do log do peso minimiza o efeito de um termo ser repetido freqüentemente.
Nenhum	$L_{ij} = a_{ij}$	O peso é a própria freqüência que o termo *i* aparece no documento *j*.

Na tabela 10, o Peso do termo considera a freqüência do termo na coleção de documentos:

Tabela 10 – Peso do Termo

Peso do Termo	*Fórmula*	*Descrição*
Entropia[8]	$G_i = 1 + \sum_j \frac{P_{ij} \log 2(P_{ij})}{\log 2(n)}$	Esse peso destaca os termos que ocorrem em poucos documentos da coleção.
GF-IDF	$G_i = \frac{g_i}{d_i}$	*Global Frequency multiplied by inverse Document Frequency*, assim como na entropia, privilegia a ocorrência em poucos documentos.
IDF	$G_i = \log_2(\frac{n}{d_i}) + 1$	*Inverse Document Frequency* enfatiza termos que ocorrem em poucos documentos da coleção.
Normal	$G_i = \frac{1}{\sqrt{\sum_j a_{ij}^2}}$	Proporção de vezes que o termo ocorre na coleção de documentos.

Os pesos de termos a seguir dependem de uma variável resposta categórica previamente identificada, isto é, os pesos são maximizados em relação à variável resposta. No caso de classificação, os documentos serão direcionados para uma classe ou outra, conforme a variável resposta categórica que distingue essas categorias.

Tabela 11 – Peso do Termo orientado à variável resposta

Peso do Termo	*Fórmula*	*Descrição*
Informação Mútua	$G_i = \max_k \left[\log \left(\frac{P(x_i, k)}{P(x_i)P(k)} \right) \right]$	Indica a proximidade da distribuição dos documentos que contém o termo com a distribuição de documentos que estão contidos na categoria.
Ganho de Informação	$G_i = -\sum_k P(k) \log(P(k) + P(i) \sum_k P(k \mid i) \log P(k \mid i) + P(\bar{i}) \sum_k P(k \mid \bar{i}) \log P(k \mid \bar{i})$	Indica a redução esperada na entropia que é causada pelo particionamento do corpus pelo termo, isto é, indica quanto a presença do termo, ou sua ausência, contribui para predizer a sua categoria.
Qui-Quadrado	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, onde f_e= freqüência esperada e f_o= freqüência observada	Indica a distância entre a freqüência esperada e freqüência observada em uma tabela de termo e categoria.

[8] O peso do termo, considerado aqui, é uma derivação do conceito introduzido por Shannon na Teoria da Informação. Para o leitor interessado, sugere-se a leitura de: Shannon, Claude E. 1948. A mathematical theory of communication. The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948.

x_i representa a variável aleatória binária para a presença do termo em um documento, enquanto k, a variável resposta categórica.

Exemplificando, apresenta-se a tabela de contingência em que a variável resposta binária *k* pode assumir os valores (sim, não). Isso significa que o documento pertence ou não a uma categoria. Ainda, x_i representa a presença, ou não, do termo *i* no documento.

Tabela 12 – Tabela de Contingência

	k=sim	***K=não***	***Total***
x_i=1	A	B	A+B
x_i=0	C	D	C+D
Total	A+C	B+D	n=A+B+C+D

Na qual, A é o número de documentos na categoria que contém o termo *i*, B é o numero de documentos que contém o termo *i* mas não pertence à categoria, C é o número de documentos na categoria que não contém o termo *i* e D é o numero de documentos que não pertence à categoria, nem contém o termo *i*. Então $P(x_i)$, $P(k)$ e $P(x_i,k)$ são:

Tabela 13 – Probabilidade de X_i e K

	k=sim	***k=não***	***Marginal***
x_i=1	$P(x_i = 1, k = sim) = \frac{A}{n}$	$P(x_i = 1, k = não) = \frac{B}{n}$	$P(x_i = 1) = \frac{(A+B)}{n}$
x_i=0	$P(x_i = 0, k = sim) = \frac{C}{n}$	$P(x_i = 0, k = não) = \frac{D}{n}$	$P(x_i = 0) = \frac{(C+D)}{n}$
Marginal	$P(k = sim) = \frac{(A+C)}{n}$	$P(k = não) = \frac{(B+D)}{n}$	

Na escrita, alguns termos destacam os aspectos semânticos de um documento mais que outros, isto é, existem termos, em determinados contextos, que são mais significativos do que outros em relação ao conceito intrínseco do texto. Em linhas gerais, os pesos utilizados ajudam distinguir termos considerados mais importantes em relação à captura de características de documentos (BAEZA-YATES; RIBEIRO-NETO, 1999).

O leitor interessado em mais detalhes pode verificar Yang e Pedersen (1997) e Woodfield (2004) que analisam distintas ponderações existentes e como elas influenciam no resultado final. Weiss et al (2005) e Manning e Schütze (1999)

apresentam os aspectos teóricos de alguns pesos descritos anteriormente e como utilizá-los.

O emprego de técnicas que utilizam pesos de ocorrência ou de termos em variadas aplicações com sucesso e demonstração prática de robustez não impede que se busque o desenvolvimento de modelos matemáticos da distribuição de termos. De tal modo, a boa compreensão da distribuição padrões de termos em um corpus contribui para sua caracterização mais precisa e na escolha ou eliminação de termos menos subjetiva.

5.4.5 Redução de dimensionalidade

Cada tipo de documento possui características, termos, mais ou menos apropriadas para descrevê-lo ou caracterizá-lo. A escolha destas características mais relevantes é determinante para a representação individualizada e sem perda de informação de documentos da coleção (WIVES, 2001).

Dado que cada termo é um atributo do documento e é utilizado para caracterizá-lo, o problema da alta dimensionalidade é típico do processo de DCT e a busca para melhor representatividade dos documentos sem perda de informação e da eficiência de processos computacionais é uma necessidade.

Independente da medida escolhida, a tabela de representação terá um número para indicar a presença do termo e o número zero para indicar a ausência dele, conforme a tabela 14

Tabela 14 – Alta Dimensionalidade – Documento por Termo

	T_1	T_2	T_3	T_4	...	T_{1200}
D_1	*1*	*1*	*0*	*0*	...	*0*
D_2	*0*	*0*	*1*	*0*	...	*1*
...	...	...	...	...	...	...
D_n	*1*	*0*	*0*	*0*	...	*0*

Constatam-se dois problemas neste tipo de representação que são:

Um número muito grande de termos - Transformando o documento em palavras dispostas nas colunas da tabela, o número de termos será certamente

elevado. Por exemplo, utilizando a frase anterior tem-se 22 termos distintos. Assim, generalizando para um documento é de se esperar que essa tabela possua centenas ou milhares de colunas representando os termos do texto.

Uma grande quantidade de zeros - Espera-se que os termos identificados em um documento não sejam os mesmos identificados em um segundo. Por exemplo, escolhendo-se um texto contendo 1.000 termos e um outro, 1.200 termos é provável que haja a interseção de uma certa quantidade destes termos que serão preenchidos com o número 1, conforme ilustrado a tabela 14, e o restante seja preenchido com zeros. Por exemplo, supondo que 400 termos são identificados nos 2 textos, logo serão assinalados com o número 1 tanto no texto A quanto no texto B. Dessa forma, 600 termos serão identificados somente no texto A e 800, somente no texto B. Estendendo esse raciocínio para uma grande quantidade de documentos teremos um tabela com uma enorme quantidade de colunas contendo zero, isto é, ausência do termo.

Toda a tarefa de processamento do dado textual tem se ocupado em padronizar os textos e otimizar seu conteúdo no sentido de trabalhar com o menor nível de ruído nos dados e conseguir uma quantidade ótima de termos que represente o conceito dos textos originais.

5.4.5.1 Lei de Zipf

Os termos encontram-se distribuídos nos textos de um corpus obedecendo a um padrão de maneira que existe uma relação da freqüência de alguns termos e suas posições em uma lista ordenada. Zipf (1949), professor de lingüística em Harvard (1902-1950), observou que essa relação se aplica em vários fenômenos humanos que chamou de O Princípio do Menor Esforço (*The Principle of the Least Effort*).

A aplicação dessa lei a uma coleção de textos consiste na contagem dos termos *f* e na sua ordenação *r* e, então, o produto da freqüência de cada termo *f* e sua ordem *r* na lista de termos é aproximadamente uma constante *k* tal que:

Equação 5

$$f * r = k$$

Por exemplo, comparando-se palavras de um texto tem-se:

Tabela 15 – Exemplo da Lei de Zipf

Palavra	*Freqüência (f)*	*Ordenação (r)*	*Constante K=f*r*
A	1000	1	1000
Informação	500	2	1000
Busca	333	3	1000
Ciência	250	4	1000
...	...	...	...
Texto	1	1000	1000

Que significa que se o termo mais freqüente se repete 1000 vezes, então o 2º termo mais freqüente se repetiria $\frac{k}{2}$ vezes que é 500, e, por conseguinte, o 3º termo seria contado $\frac{k}{3}$ que totaliza 333 e assim por diante.

Portanto, trata-se de contagem de palavras distribuídas nos documentos que serão objetos de estudo e a sua partição merece detalhamento para compreendê-la. Dessa forma, a lei de Zipf é uma constatação empírica e apresenta uma descrição da distribuição de freqüências de palavras na linguagem humana: existem poucos termos muito comuns, uma quantidade média de termos de freqüência intermediária e muitos termos que ocorrem poucas vezes, conforme figura 4:

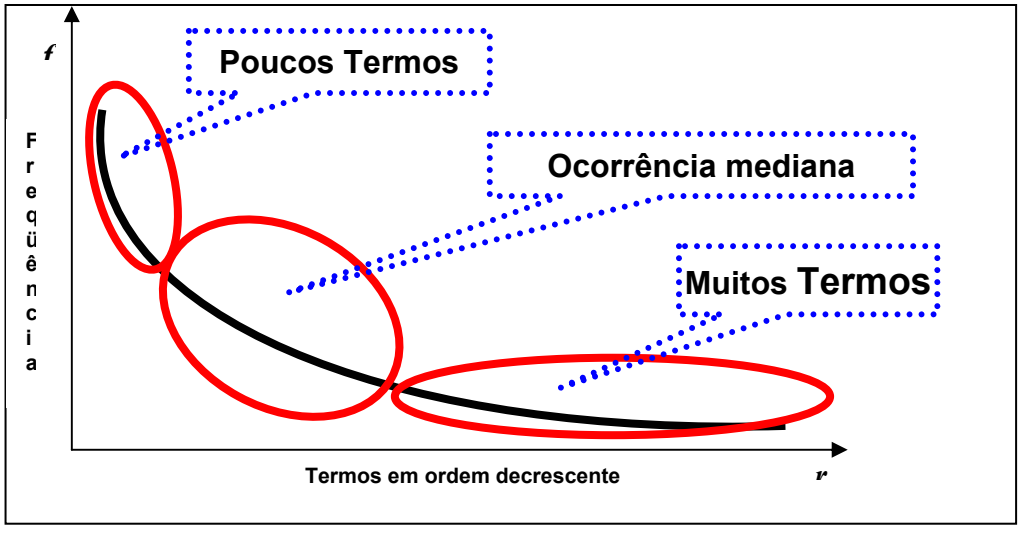

Figura 4 – Representação Gráfica da Lei de Zipf

Essa figura pode auxiliar seleção de pontos de cortes para remoção de palavras com baixo poder de discriminação de documentos e, assim, contribui para a redução de dimensionalidade. Ressalta-se que a escolha do ponto de corte é feita de maneira arbitrária. Deve-se levar em conta a experiência do analista para que se mantenha o menor número de termos aliada a menor perda de informação.

5.4.5.2 Significância das Palavras de Luhn

Luhn (1958) em seu artigo *The Automatic Creation of Literature Abstract* propõe a identificação automática de tópicos em artigos com o propósito de criar resumos automáticos. O conceito vem ao encontro do propósito da DCT, pois a identificação de pontos chaves no texto para capturar a idéia de escritores é um dos processos requeridos para o processo de descoberta.

A divisão de textos em capítulos, parágrafos, orações, frases, etc são manifestações físicas da associação de idéias do escritor. Assim, na linguagem escrita, as idéias mais associadas intelectualmente são implementadas por palavras mais associadas fisicamente (LUHN, 1958).

Um escritor normalmente repete palavras à medida que avança ou varia sua argumentação e assim elabora os aspectos de seu assunto. A freqüência de um termo em um documento fornece uma medida útil para determinar a significância de uma palavra (LUHN, 1958). Essa abordagem não leva em conta as relações lógicas ou semânticas do escritor.

Segundo Moens (2000), foi Luhn quem descobriu que padrões de distribuição de termos poderiam fornecer informação significativa sobre o conteúdo de um documento. Altas freqüências de termos tendem a ser comuns e não são relevantes para destacar o conteúdo. Por outro lado, uma ou duas ocorrências de um termo em textos relativamente longos também podem não fornecer informação relevante na descoberta do assunto apresentado no documento.

Utiliza-se a abordagem de Zipf que cria uma lista de termos em ordem decrescente de freqüência e, então, identifica a sua relevância em função do

assunto do documento. A idéia de Luhn é de que existem pontos de corte que podem ser calculados através de métodos estatísticos ou atribuídos pela experiência de analistas de domínio. Esses pontos delimitam as ocorrências dos termos que são significativos para a identificação do tema.

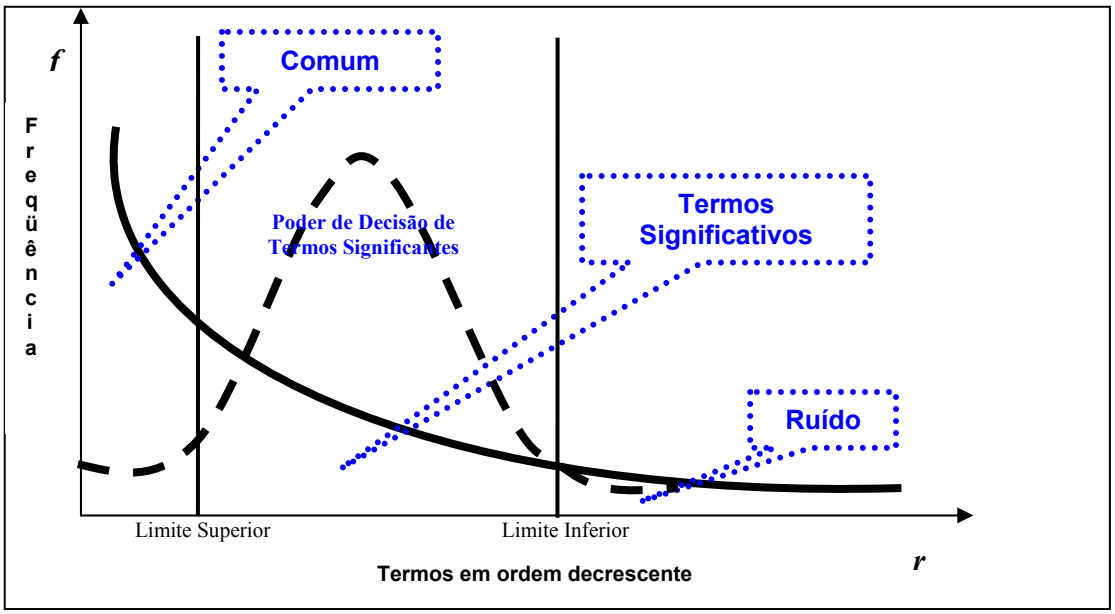

Figura 5 – Significância de Termos - Adaptação de Luhn (1958), pág. 11

Conforme figura 5, propõe-se que os termos à esquerda do limite superior são comuns e aqueles à direita do limite inferior são raros e, portanto, não contribuem significativamente para o conteúdo do texto. Assim, os termos que contribuem significantemente estão entre os limites superior e inferior. Ainda, existe uma curva que Luhn chamou de poder de decisão de termos significantes, que expressam a capacidade de discriminar o conteúdo, ilustrando que os termos, em uma ordem de significância imaginária que se inicia próxima de zero, vão crescendo em habilidade de discriminação até atingirem o pico na metade entre os limites superior e inferior e então começam a diminuir simetricamente até o último termo.

Certa arbitrariedade está envolvida na escolha destes limites. Não há oráculo que forneça esses valores e eles tendem a ser estabelecidos por tentativa e erro. Destaca-se que essa análise não se aplica somente aos termos, mas também às frases e termos lematizados. (VAN RIJSBERGEN, 1979).

Verifica-se a relação entre a curva de Zipf e o conceito de Luhn na identificação de onde os termos significantes estão, pois ambas apontam os termos de baixa significância nas extremidades da distribuição dos termos (MOENS, 2000).

Um dos problemas inerentes aos descartes de termos com o objetivo de reduzir a dimensionalidade é a perda da informação. O ideal é que vários termos sejam combinados em um só mantendo a informação original com uma dimensão menor. A conversão de termos sinônimos para o termo preferido é um bom exemplo de redução de dimensionalidade sem perda de informação.

Decomposição de Valores Singulares - DVS[9]

Segundo Manning e Schütze (1999), as técnicas de redução de dimensionalidade extraem um grupo de objetos que existem no espaço com muitas dimensões e os representa no espaço com poucas, em geral, duas ou três dimensões com a finalidade de visualização.

O modelo de espaço vetorial é uma representação freqüente na recuperação de informação principalmente pela sua simplicidade conceitual e utilização de proximidade espacial para denotar similaridade semântica entre documentos.

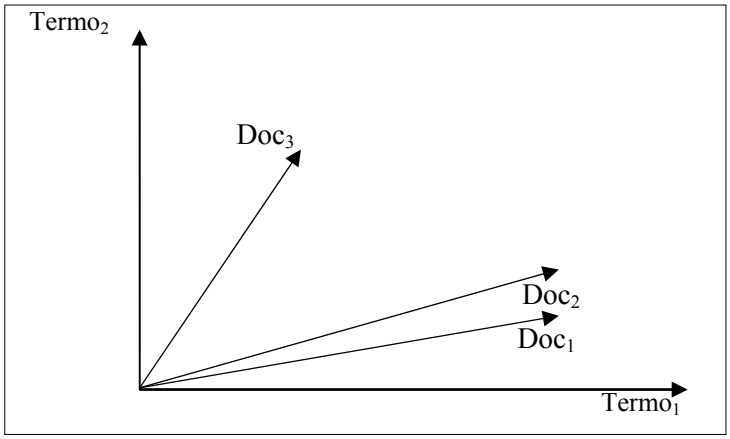

Figura 6 – Espaço Vetorial em duas dimensões

Esse modelo é uma maneira de representar documentos espacialmente por meio das palavras que ele possui. Cada documento é representado na forma de um vetor em relação ao vocabulário que contém.

[9] do inglês Singular Value Decomposition - SVD

A figura 6 mostra a representação em duas dimensões correspondentes aos termos 1 e 2 e três documentos (1, 2 e 3) no espaço. A proximidade de vetores é calculada pelo ângulo, isto é, quanto menor o ângulo entre dois vetores, mais próximos semanticamente eles são. Nesse exemplo, os documentos 1 e 2 possuem uma proximidade espacial muito maior que com o documento 3, isto indicaria que os documentos 1 e 2 são mais similares semanticamente.

O espaço vetorial original é constituído de termos únicos que ocorrem nos documentos e, mesmo em uma coleção de textos de tamanho moderado, eles podem chegar a dezenas ou centenas de milhares. Entretanto, para grande parte de algoritmos de mineração, isto é um fator proibitivo. Dessa forma, deseja-se a redução de dimensionalidade sem perda de informação (YANG; PEDERSEN, 1997).

Freqüentemente, por questões práticas, utiliza-se a matriz de representação dos dados da forma termo-documento como se segue:

Tabela 16 – Matriz Termo-Documento

T_1	*1*	*1*	*0*	...	*0*
T_2	*0*	*0*	*0*	...	*1*
T_3	*1*	*0*	*0*	...	*1*
T_4	*0*	*0*	*1*	...	*1*
...	...	...	...	...	...
T_n	*1*	*0*	*0*	...	*0*

A tabela 16 indica n dimensões de termos distribuídos através de m documentos de tal forma que a meta será extrair dimensões que contenham relações semânticas entre os documentos, isto é, a combinação de termos que componham conceitos relacionados aos documentos

Desse modo, a Decomposição em Valores Singulares, ou simplesmente DVS, é uma técnica matemática de redução de dimensionalidade que visa formar novas variáveis que são combinações lineares das variáveis originais. A finalidade é utilizar um número muito menor de novas variáveis que contêm a informação das variáveis originais, isto é, empregam-se poucas variáveis sem perda de informação.

Formalmente, tem-se um espaço n-dimensional (n termos) que é projetado sobre um espaço k-dimensional, onde $k < n$. O objetivo é formar novas variáveis que

são combinações lineares das variáveis originais. (SHARMA, 1996). Deve-se observar que pragmaticamente *k* deve ser muito menor que *n*, caso contrário, não haveria benefício no emprego da técnica.

Portanto, a projeção transforma um vetor de documentos no espaço n-dimensional de termos para um vetor em um espaço k-dimensional reduzido. Como ilustração, é equivalente a representar geometricamente duas variáveis originais em apenas uma, ou seja, DVS é muito semelhante a ajustar uma reta, um objeto uni-dimensional, a um conjunto de observações que existe no plano (bi-dimensional) (MANNING; SCHÜTZE, 1999).

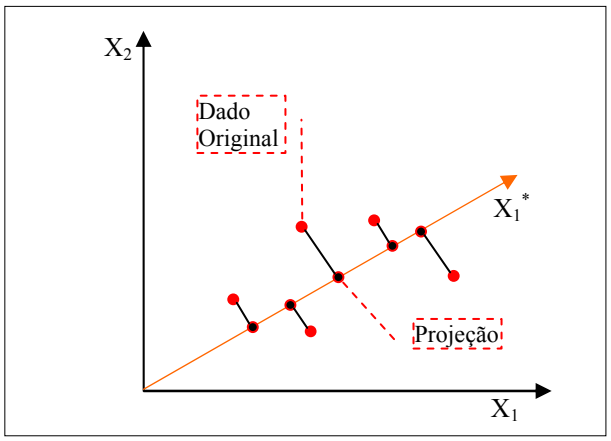

Figura 7 – Projeção de observações.
Adaptação de Sharma (1996), pág. 65.

A seleção de projeções em dimensão menor é feita geralmente pela otimização de características interessantes nos dados originais através de todas as direções de projeções, isto é, a captura da máxima variação dos dados em uma quantidade menor de variáveis (KHATTREE; NAIK, 2000).

A DVS é computada pela decomposição de uma matriz A que contenha os vetores de documentos com cada coluna correspondente a um documento, isto é, o elemento a_{ij} da matriz registra a freqüência do termo ***i*** no documento ***j***. Quaisquer dos pesos discutidos anteriormente podem ser utilizados para popular a matriz.

Segundo Sharma (1996), Manning e Schütze (1999), Khattree e Naik (2000) e WoodField (2002), a matriz A é decomposta no produto entre das matrizes T, S e D.

Equação 6

$$A_{txd} = T_{txn} * S_{nxn} * (D_{dxn})^t ,$$

onde t = quantidade de termos,
d = quantidade de documentos,
n = min(t, d),
D^t = é a transposta de D

Cada matriz T e D é ortogonal, que significa que a matriz T^t *T = I e D^t *D = I. S é uma diagonal.

A DVS pode ser vista como um método para percorrer os eixos do espaço n-dimensional tal que o primeiro eixo percorre a maior variação entre os documentos, a segunda dimensão percorre a dimensão com a segunda maior variação e assim por diante. O número máximo de dimensões é a quantidade de termos da matriz A.

As matrizes T e D representam os termos e documentos, respectivamente, neste novo espaço. A diagonal de S contém os valores singulares de A em ordem decrescente de maneira que o i-ésimo valor singular corresponde à quantidade de variação ao longo do i-ésimo eixo.

A projeção da DVS é a combinação linear das linhas e termos da matriz termo-documento original. A idéia de combinação linear pode ser entendida como uma extensão de média ponderada dos termos. Essa média ponderada produz o conceito contido nos documentos e, daqui, considera-se que a DVS converte termos em conceitos.

Matematicamente, essa projeção forma subespaço k-dimensional que representa o melhor ajuste para descrever os dados originais. A projeção de colunas da matriz termo-documento é um método para representar cada documento por k conceitos distintos. Em outras palavras, a coleção de documentos é mapeada no espaço k-dimensional no qual cada dimensão é reservada para cada conceito. Da mesma forma, cada linha, ou termo, pode ser projetada sobre as k primeiras colunas de S. Enfim, a técnica da DVS encontra a projeção ótima para um espaço reduzido,

de maneira que representa os termos e documentos da melhor forma possível em um espaço dimensional menor.

Finalmente, a aplicação da DVS na área de Recuperação da Informação é chamada de Indexação por Semântica Latente, ou usualmente em inglês, *Latent Semantic Indexing*. Essas novas dimensões são uma melhor representação de documentos e de consultas. O nome “latente” é uma metáfora devido ao fato de que essas novas dimensões são a representação verdadeira, pois a ISL recupera a estrutura semântica original do espaço e suas dimensões originais.

5.5 Considerações Finais

A descoberta de conhecimento em textos é a conjunção de várias metodologias e conceitos, logo esse trabalho apresenta uma reprodução estática de seu desenvolvimento e implementações até esse momento. O desenvolvimento e aperfeiçoamento do processo são constantes dada a natureza da língua e das ferramentas tecnológicas.

Konchady (2006) declara que a DCT é uma prática relativamente nova derivada da Recuperação da Informação – RI – e da PLN e essa afirmação estabelece um vínculo significativo, visto que a RI é uma das áreas principais de pesquisa da CI.

Segundo Bräscher (1999), os avanços tecnológicos influenciam a CI e favorecem o surgimento de novas técnicas de representação e recuperação de assunto considerando os aspectos cognitivos envolvidos no processo de comunicação homem-máquina que exigem modelos de representação do conhecimento capazes de contextualizar os significados expressos nos textos armazenados.

Lima (2003), em uma releitura de Saracevic (1995), expõe que a CI é uma área interdisciplinar que reúne a Biblioteconomia, a Ciência Cognitiva, a Ciência da Computação (CC) e a Comunicação, com forte associação dos processos humanos da comunicação e da tecnologia no seu contexto contemporâneo

De fato, a CC trata de algoritmos relacionados à informação, enquanto a CI se dedica a compreensão da natureza da informação e de seu uso pelos humanos. A

CI e a CC são áreas complementares que conduzem a aplicações diversas (SARACEVIC, 1995).

Robredo (2003) reafirma a interdisciplinaridade da CI orientando que não se pode restringir o escopo e a abrangência da informação ao campo exclusivo da biblioteconomia e da CI, pois variados estudiosos, pesquisadores e especialistas lidam com a informação de um ponto de vista cientifico e nas mais variadas abordagens e aplicações. Ainda, ensina que ela pode ser dividida, para fins de estudo e delimitação do(s) objeto(s), mas sem perder de vista o interesse comum de todos os seus domínios, a entidade informação.

Ainda, Lima (2003) aponta as possibilidades de interseção entre a CI e a CC que se concentram nos processos de categorização, indexação, recuperação da informação e interação homem-computador.

Nesse sentido, a DCT pode ser vista como a interposição da Estatística que utiliza métodos quantitativos para transformar dados em informação, da CC fornece suporte tecnológico para manipulação dela e da CI que concentra o foco de atuação na sua gestão.

Diante do exposto, a DCT encontra-se nesta área de intersecção da CI, da Estatística e da CC que utiliza métodos lingüísticos para tratamento de textos. Essas áreas são mutuamente beneficiadas pelo aporte teórico de cada uma que favorece o desenvolvimento conceitual interdisciplinar. De tal modo, fica caracterizada a evidente contribuição do estudo da DCT no âmbito da CI.

6 METODOLOGIA

O trabalho consiste na descoberta de conhecimento nos textos contidos na base do SAC. Nesta base, o campo com a argumentação em linguagem natural fornecida por clientes é o texto a ser minerado.

O tema da pesquisa na área da Ciência da Informação é muito pouco explorado, portanto vem promover a discussão que permanece quase exclusivamente no domínio da Ciência da Computação. Além disso, no âmbito da Instituição Financeira em que será aplicado, caracteriza-se pelo ineditismo e conta com o apoio da área responsável pela gestão do sistema de SAC.

6.1 Recursos Utilizados

Os recursos utilizados para a execução do projeto são:

Programa

- Microsoft Windows XP Professional service pack 2 – sistema operacional;
- Microsoft Excel 2002 – manipulação de dados e criação de gráficos;
- SAS 9.1.3 service pack 4 – programação e manipulação de dados;
- SAS Enterprise Miner 4.3 – interface gráfica utilizada para mineração e modelagem de dados;
- SAS Text Miner 3.1 – ferramenta de mineração de textos.

Equipamento

- PC AMD Athlon™ XP 2600+ com 512 Mb de memória RAM e 40 Gb de disco.

Dados

- Base de dados do Serviço de Atendimento ao Consumidor desde sua criação em agosto de 2000 até junho de 2006;

- Cada registro compõe uma mensagem completa;
- O arquivo foi disponibilizado em formato de texto para que pudesse ser importado e manipulado pelo programa de computador utilizado.

6.2 O Programa SAS

A escolha do programa SAS deveu-se à prática de quase dez anos de experiência na utilização da ferramenta que possibilitaria a imediata utilização de seu potencial. Além disso, o conhecimento dos recursos disponíveis que foram fundamentais para o desenvolvimento da pesquisa no prazo que uma pesquisa deste tipo requer.

A ferramenta utilizada para todo processo é o SAS Enterprise Miner que é capaz de fornecer uma interface amigável para a construção de modelos e, além disso, uma eficiente linguagem de programação para resolução de problemas específicos. O programa suporta a leitura de textos em vários formatos como por exemplo, word, html, pdf, ASCII entre outros. Possui ferramentas para pré-processar o texto e dispositivos lingüísticos para executar a lematização. Ressalta-se que o software não tem o mecanismo de rotulação de partes do discurso, nem realiza a extração de entidades disponíveis em português.

O programa possui, ainda, ferramentas de visualização e de exploração destinadas à análise exploratória do corpus e algoritmos de transformação de variáveis como a DVS e de agrupamentos de documentos. Para a criação de modelos de classificação conta com ferramentas especializadas de modelagem.

Os modelos utilizados neste estudo são:

- Redes Neurais - Sistemas que inicialmente tentavam imitar a neurofisiologia do cérebro humano através da combinação de construções matemáticas relativamente simples (neurônios) em um sistema fortemente conectado utilizando análise numérica e estatística;
- Regressão Logística - Técnica estatística que investiga o relacionamento entre uma variável dependente categórica e um grupo de variáveis independentes categóricas ou contínuas;

- Árvore de Decisão - Consiste na segmentação dos dados pela a aplicação de uma série de regras que buscam maximizar as diferenças na variável dependente;
- Memory-based Reasoning – processo que identifica casos similares e aplica as informações obtidas destes casos aos novos casos.

Ao leitor interessado nos aspectos teóricos na construção desses modelos e na aplicação destas técnicas no mundo dos negócios, sugere-se a leitura de Berry & Linoff (1997).

Para validação dos modelos, existem procedimentos automatizados que quantificam os resultados encontrados para que o analista possa fazer inferências.

Nesse sentido, foi escolhido aquele que fez a melhor categorização baseada na extração de três amostras aleatórias para treinamento, teste e validação do modelo. Essas amostras são comparadas e o próprio programa fornece indicadores do desempenho do categorizador.

6.3 Tipo de Pesquisa

A pesquisa é do tipo descritiva, pois o escopo é extrair características da população, no caso as mensagens do SAC, e estabelecer correspondência entre variáveis e definir sua natureza. Visa ainda descrever o processo da Organização, atendimento ao cliente, utilizando métodos quantitativos e amostragem que caracterizam a abordagem quantitativa.

6.4 Método de abordagem

O trabalho foi baseado em amostra do sistema SAC, na revisão da literatura que trata, principalmente, da utilização de técnicas de descoberta de conhecimento e estudos de clientes sob a perspectiva do Serviço de Atendimento ao Consumidor.

Verifica-se a necessidade de se otimizar o uso da base do SAC com automatização de processos. Portanto, trata-se de um levantamento com mais profundidade de um caso específico, o que caracteriza um estudo de caso.

Tal estudo se faz necessário em função do potencial recurso de conhecimento explícito, a ser explorado, através das manifestações por escrito dos clientes que podem ser transformadas em fonte preciosa de informação estratégica.

6.5 Fluxo Operacional

Na figura 8, encontra-se ilustrado o processo global de desenvolvimento da pesquisa.

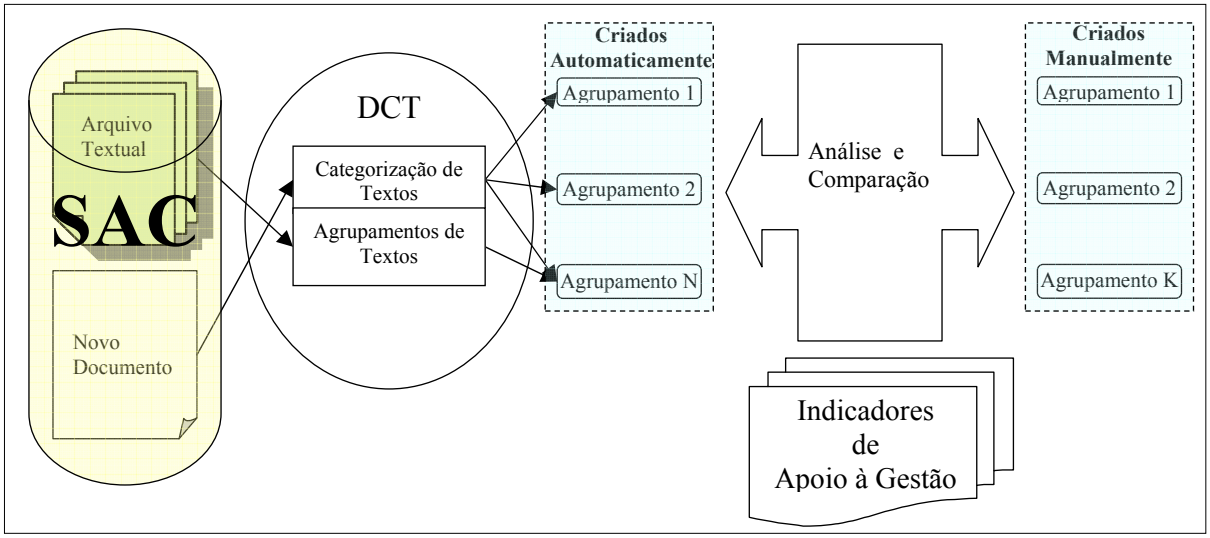

Figura 8 – Processo de desenvolvimento da pesquisa

A base SAC contém os documentos necessários para o estudo. Os arquivos textuais históricos são o insumo para as técnicas de mineração de texto que produzem agrupamentos cujo objetivo é reunir os documentos similares. Após essa etapa, esses grupos criados por processo automático serão submetidos aos analistas de domínio para validação e interpretação de seu conteúdo. Em seguida, esses agrupamentos serão comparados aos já existentes, criados manualmente.

Uma vez estabelecidos os grupos que representam de forma substantiva a categoria dos documentos utiliza-se essa nova variável para a criação de um modelo categorizador de novos documentos.

Finalmente, tem-se a base textual segmentada e estruturada que possibilita a extração de relatórios que fornece indicador para apoiar a gestão.

6.6 Utilização da DCT

Após a coleta e seleção dos documentos, um passo essencial e que consome a maior parte do tempo é a preparação dos dados. O processo envolve várias tarefas que, muitas vezes, são relacionadas à área de pesquisa e, para tanto, necessita-se do apoio de analistas de domínio.

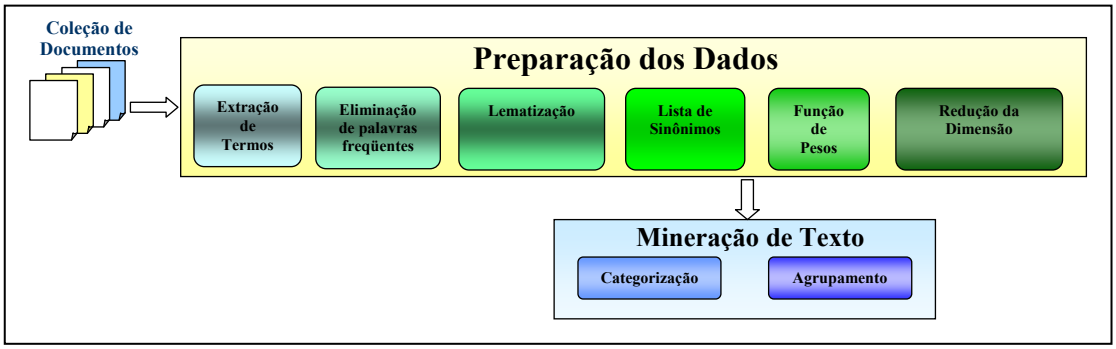

Figura 9 – A DCT da pesquisa

A figura 9 apresenta o detalhamento da preparação de dados, bem como os dois algoritmos utilizados para identificação de padrões nos dados:

A extração de termos é executada automaticamente e normalmente o espaço em branco entre as palavras é o indicador para dividir o texto em termos. Deve-se ter o cuidado de não separar palavras compostas, como por exemplo, “mineração de textos” e por isso faz-se necessária a execução de testes estatísticos para verificar e validar a co-ocorrência entre termos e a criação de dicionário contendo as palavras compostas.

A eliminação de palavras freqüentes já é um processo com certa estabilidade nos programas especialistas e pode ser feito de maneira automática, tomando-se o cuidado apenas de verificar se os termos escolhidos se distribuem igualmente por toda a coleção para o descarte.

Alguns estudos já demonstraram que a lematização pode incorrer em erro de redução de termos com sentidos diferentes, porém com mesmo radical. Entretanto, esse erro é de aproximadamente 5%, o que não deve alterar o resultado final.

A criação da lista de sinônimos foi feita em parceria com analistas de domínio e uso de dicionário especializado. Dessa forma, extraíram-se, segundo seus pesos,

os termos mais relevantes que foram validados junto aos especialistas e manualmente foram alimentadas as suas formas sinônimas. Por outro lado, foi criado um programa que verifica as palavras que se repetiam menos de 6 vezes em toda a coleção e comparada a um dicionário[10] do Português Brasileiro para corrigir o termo quando fosse o caso ou confirmar o termo pesquisado. Nesse dicionário foi também possível resolver formas verbais não identificadas pelo lematizador da ferramenta. Dessa forma, foi criado um dicionário especializado de sinônimos ao conjunto de termos.

Nesta etapa, foram consideradas mais de uma função de pesos disponível no software utilizado. O objetivo foi verificar o melhor ajuste e efetiva utilização.

A redução de dimensionalidade foi feita com a eliminação de termos que utilizou alguns processos descritos à frente e a utilização da transformação DVS que também objetiva a redução da quantidade de termos.

6.6.1 O Corpus e a Preparação de Dados

Os procedimentos descritos a seguir, tratam da descrição da base de dados e sua efetiva preparação para utilização na mineração de textos.

6.6.1.1 A Base de Dados

A base explorada foi extraída do sistema de ouvidoria que armazena todas as mensagens desde 2000. Essa base é atualizada diariamente, portanto, para a produção de material para esse estudo estabeleceu-se que o corte deveria ser até o mês de junho de 2006 inclusive. A organização dos campos analisados da base do SAC é apresentada na tabela 17.

[10] Dicionário Unitex-PB disponível em http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html

Tabela 17 – Descrição da Base

Grupo Assunto *– Assunto a que se refere a mensagem;*
Origem – o canal por onde se fez o contato (Internet, telefone,...);
Natureza – se reclamação, sugestão ou elogio;
Produto – o produto do qual se está comentando;
Motivo – pré-classificação do atendente;
Descrição da Ocorrência – Texto descrevendo a ocorrência;

O exemplo ilustrado na tabela 18 mostra um registro extraído da base estudada.

Tabela 18 – Exemplo da Base de Dados

GRUPO ASSUNTO	***ORIGEM***	***NATUREZA***	***PRODUTO***	***MOTIVO***	***DESCRIÇÃO DA OCORRÊNCIA***
AGENCIA	TELEFONE	RECLAMAÇÃO	AGÊNCIA	FALTA DE CORTESIA DO GERENTE	Cliente reclama que dia 07/02 esteve nesta agência para solicitar um empréstimo e a Gerente Sra.X informou os documentos necessários. Dia 08/02 retornou a agência e ao entregar o RG a Sra.X ela dirigiu-se ao Gerente Sr.Y com irônia e disse que o documento não seria aceito pelo caixa. A cliente foi falar com o Gerente Sr.Y que não aceitou o RG, mesmo a cliente dizendo que não houve tempo dísponível para retirar outro, mas o gerente mencionou em voz alta que o RG estava um "lixo" e a cliente era incompetente até mesmo para retirar outro documento. Pede providências urgentes quanto ao atendimento prestado.

6.6.1.2 Descrição da Base de Dados

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

Nesta fase, o apoio de especialistas de domínio foi fundamental para que se pudessem traduzir os resultados obtidos em informações que fossem de interesse do gestor do sistema.

6.6.1.3 Definição do escopo do Corpus

Segundo a opinião de especialistas, esse tipo de informação deve estar tão atualizada quanto possível, pois se tratam de problemas pontuais apontados pelos clientes que encontram dificuldades ou insatisfação em relação a produtos ou serviços da empresa naquela data específica. Esse apontamento pode variar com o passar do tempo.

Inicia-se a descrição observando a distribuição dos registros em relação ao tempo. A tabela abaixo representa todo o arquivo até junho de 2006, isto é, desde a sua criação.

Tabela 19 – Distribuição Anual das Ocorrências

Data do Chamado	***Freqüência***	***%***	***Freqüência Acumulada***	***% Acumulado***
2000	6.292	1,26	6.292	1,26
2001	47.454	9,51	53.746	10,77
2002	87.080	7,45	140.826	28,22
2003	94.292	18,89	235.118	47,11
2004	75.178	15,06	310.296	62,17
2005	105.202	21,08	415.498	83,25
2006	83.604	16,75	499.102	100,00

De acordo com a tabela 19, existem 499.102 registros. No último ano, 2006, o número de registros corresponde a 6 meses. Embora tenha havido uma diminuição em 2004 em relação a 2003, a utilização do sistema mostra uma tendência de crescimento indicando que o cliente vem se acostumando com um canal no qual pode manifestar sua opinião à empresa.

A figura 10 mostra um gráfico que destaca essa tendência de crescimento de maneira mais clara:

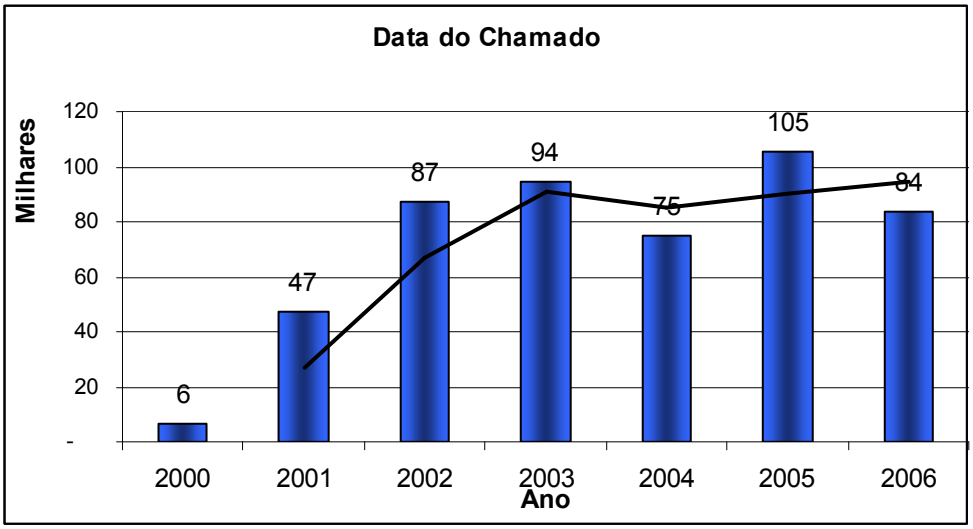

Figura 10 – Evolução Anual de Registros

Considerando o volume de dados e a relevância de informações atualizadas espera-se que as informações mais recentes retratem mais fidedignamente a realidade atual, portanto a base para o estudo se restringirá ao ano de 2006 que contém os meses de janeiro a junho.

Continuando a delimitação do escopo do projeto, descobrir o quê incomoda o cliente é uma meta pretendida por qualquer empresa que se preocupa em conquistar o mercado e assegurar um lugar no topo da preferência. Para tanto, parece óbvio que concentrar os esforços de estudo na reclamação feita pelo cliente é algo natural.

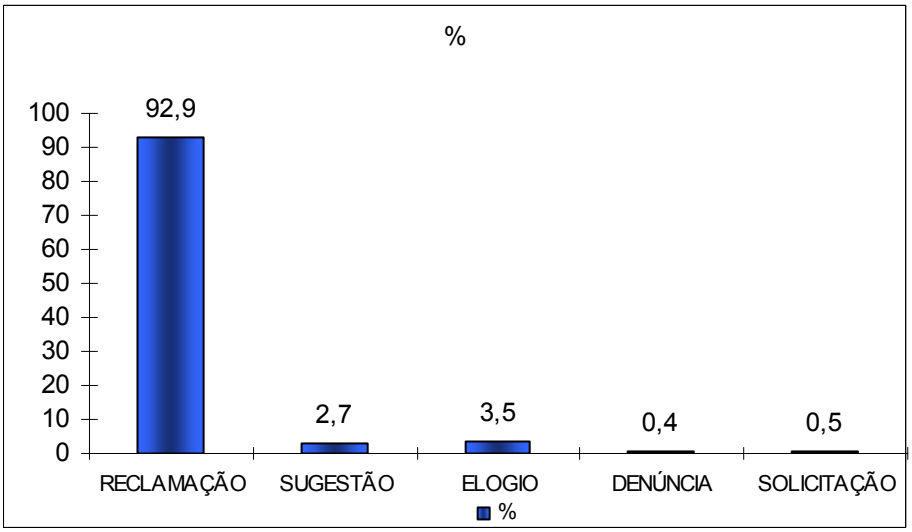

Figura 11 – Natureza das Ocorrências

Apesar de outros tópicos constantes da base de estudo serem importantes a pesquisa aprofundou-se no tema reclamação de acordo com necessidade mais premente da empresa. De fato, observa-se na figura 11 a preponderância do tópico “Reclamação” em 92,9% dos casos constantes da base. Isto significa um enorme potencial de estudo e possibilidades de fidelização do cliente. Daqui o estudo reduziu o seu escopo para o ano de 2006 e para o item Reclamação.

Um outro ponto de interesse é identificar o canal que concentra a entrada de dados e, ainda, que ele seja padronizado, visando facilitar a descoberta de conhecimento com a utilização da mineração de textos.

Tabela 20 – Canal de Origem das Ocorrências

Origem do Contato	Freqüência	%	Freqüência Acumulada	% Acumulado
TELEFONE	52.646	67,81	52.646	67,81
OUVIDORIA	11	0,01	52.657	67,83
IMPRENSA	4	0,01	52.661	67,83
PAPELETA	130	0,17	52.791	68,00
PROCON – AUD.	28	0,04	52.819	68,04
C. C. DO FGTS	1	0,00	52.820	68,04
INTERNET	23.516	30,29	76.336	98,33
CARTA	2	0,00	76.338	98,33
BACEN	1.016	1,31	77.354	99,64
PROCON – CIP	278	0,36	77.632	100,00

Segundo observa-se na tabela 20, a preferência do cliente ao utilizar o SAC é o telefone, seguido pela Internet. Considerando que a reclamação via telefone sempre passa por um atendente, é razoável supor que a informação transcrita passe por uma padronização em conseqüência do treinamento oferecido aos trabalhadores deste local. Portanto, o estudo se limitará somente a esse tópico.

Dessa forma o arquivo fica delimitado ao primeiro semestre de 2006, registros referentes à Reclamação e originados via Telefone. O que caracteriza uma redução do universo para 52.646 mensagens textuais.

6.6.1.4 Preparação do Corpus

Um dos maiores esforços na mineração de textos é a preparação da base de dados. Esse trabalho pode ser desde a retirada de registros duplicados, até a eliminação de caracteres estranhos como se poderá ver à frente.

Verifica-se que existem 1.370 registros supostamente duplicados, entretanto, eliminá-los prontamente pode retirar informação importante. Diante disto, verificou-se registro a registro e constatou-se algumas duplicações que foram eliminadas. Conforme tabela abaixo:

Tabela 21 – Exemplo de Duplicação de Registro

Produto	Motivo	Descrição da Ocorrência	Data do Chamado	Hora do Chamado	reg
5	2076	A cliente relata que tenta entrar em contato na Central de Atendimento no 99999999 para fazer um agendamento na agência 9999 para dar entrada no seu FGTS e não consegue, pois todos os operadores estão ocupados. Pede providência urgente.	28ABR2006	05:46:35	342
5	2076	A cliente relata que tenta entrar em contato na Central de Atendimento no 99999999 para fazer um agendamento na agência 9999 para dar entrada no seu FGTS e não consegue, pois todos os operadores estão ocupados. Pede providência urgente.	28ABR2006	05:47:05	343

Os registros apresentados na tabela 21 indicam uma duplicação, pois o único campo diferente é a Hora do Chamado e essa diferença é de 30 segundos. Nesse intervalo de tempo seria impossível o atendente receber e transcrever esse chamado. O campo **reg** não é levado em conta, pois é automático e indica a quantidade de registros. Nessas circunstâncias, a eliminação de registro é segura. Foram eliminados 629 registros nessas condições restando 741 para análise mais detalhada.

Os registros com as datas e os horários diferentes podem ser transcrições freqüentes que os operadores anotam de maneira padronizada ou utilizam a facilidade do Windows de copiar e colar. Entretanto, não há como ter certeza de que se trata de uma duplicação e, portanto, esse tipo de registro não foi eliminado. A situação é representada na tabela 22.:

Tabela 22 – Exemplo de Possível Duplicação de Registro

Produto	Motivo	Descrição da Ocorrência	Data do Chamado	Hora do Chamado	reg
83	111	A cliente relata que abriu uma CONTA POUPANÇA há mais de um mês e até o momento o CARTÃO não foi enviado. Pede providência urgente.	18MAI2006	08:21:48	573
83	111	A cliente relata que abriu uma CONTA POUPANÇA há mais de um mês e até o momento o CARTÃO não foi enviado. Pede providência urgente.	28ABR2006	06:19:20	572
83	111	A cliente relata que abriu uma CONTA POUPANÇA há mais de um mês e até o momento o CARTÃO não foi enviado. Pede providência urgente.	11ABR2006	01:40:51	571
83	111	A cliente relata que abriu uma CONTA POUPANÇA há mais de um mês e até o momento o CARTÃO não foi enviado. Pede providência urgente.	03ABR2006	08:31:14	570

Nesta análise, observam-se também registros com a ocorrência igual, porém com o código do produto ou com código do motivo divergentes. Tal fato caracteriza o erro de classificação dos operadores. Segue o exemplo:

Tabela 23 – Exemplo de Duplicação de Registro e Erro de Classificação

Produto	Motivo	Descrição da Ocorrência	Data do Chamado	Hora do Chamado	Reg
183	1838	Cliente relata que abriu uma conta a mais de um mês e até a presente data não recebeu o Cartão. Pede providências.	28MAR2006	09:28:58	15377
83	111	Cliente relata que abriu uma conta a mais de um mês e até a presente data não recebeu o Cartão. Pede providências.	28MAR2006	09:07:56	15376

Na análise mais detalhada, não há segurança em afirmar que não sejam erros de operadores e estejam realmente duplicados, apesar de a data ser idêntica e de o horário apresentar diferença inferior a uma hora. Considerando a quantidade de registros e o objetivo de criação de agrupamentos, ponderou-se que a exclusão das repetições traria mais benefícios à execução das tarefas posteriores do que deixá-los e adicionar o viés advindo da provável duplicação. Foram eliminados 313 registros nestas condições.

O total de registros duplicados que foram eliminados é 942 e, portanto, o arquivo sem duplicações possui 51.704 registros.

6.6.1.5 Exploração do Texto

Um outro passo foi a verificação no texto à procura de possíveis problemas que poderiam afetar a contagem dos termos.

Inicialmente a coleção de textos possui 2.413.586 termos em 51.704 documentos. Desses termos, foram encontrados 26.337 termos distintos, isto significa dizer que o vocabulário da coleção de documentos em estudo limita-se a essa quantidade de termos encontrados.

Tabela 24 – Palavras por Documento

Quantidade	Média	Desvio Padrão	Mínimo	Máximo
51.704	46,68	23,46	1	353

Têm-se, em média, 46,7 termos por documento. Observa-se o desvio padrão alto em relação à média que indica uma grande variação no tamanho dos documentos. A informação é confirmada pela quantidade mínima e máxima de documentos que varia de 1 até 353 termos por documento. Essa observação era esperada, uma vez que algumas pessoas são prolixas na sua reclamação enquanto outras são muito sucintas.

Na tabela 24 o número mínimo de termos por documento correspondente a apenas um termo parece indicar um erro. Uma verificação mais detalhada foi realizada e, para tanto, selecionaram-se documentos com menos de 10 termos para validá-los ou excluí-los.

Tabela 25 – Documentos com menos de 10 termos

Documento	*Descrição da Ocorrência*	*Quantidade de Termos*
4.556	Cliente contesta os rendimentos do PIS .	6
11.904	Cliente relata q	3
46.138	Demora no atendimento.	3
46.144	Diante dos fatos pede providências.	5
51.294	RECLAMA QUE NÃO RECEBE O BOLETO	6
51.425	Teste	1
51.600	conta 19033-8	1
51.693	Test	1
51.694	Teste	1
51.695	Teste de envio	3
51.696	Teste de envio para a cetel, descartar...MARCOS - GEOUV	9
51.697	Teste, favor cancelar	3
51.700	X	1

Percebe-se claramente que apenas os documentos hachurados na tabela 25 (4.556, 46.138 e 51.294) são válidos, enquanto os outros são testes ou informações sem significados e, portanto, foram descartados. O corpus passou para 51.694 documentos.

Um problema muito comum na importação de arquivos texto é a identificação de caracteres estranhos que modificam o tamanho das palavras e induzem ao erro na sua identificação por algoritmos computacionais. Utilizando a programação para transformar todo texto em letras podem-se visualizar caracteres que não são usuais na escrita e que, provavelmente, são erros ortográficos ou problemas na importação do arquivo.

No exemplo a seguir, os termos ‘a’ e ‘cobrança’ aparecem como se fossem somente um, pois existe um caractere não visível, mas que os algoritmos interpretam como integrante do termo ‘a_cobrança’. Um tratamento específico para tal problema foi realizado, a fim de garantir que o processo de separação dos termos fosse o melhor possível. A tabela 26 apresenta exemplos:

Tabela 26 – Termos com Caracteres estranhos

termo	Identificador do termo	Identificador do Documento
a cobrança	74	51.275
a suspensão	19	51.275

Foram alterados 26,62% dos documentos que possuíam algum caractere estranho, isto é, 13.759 de 51.694.

Após essas alterações, efetuou-se a contagem dos termos novamente com 2.416.285 termos e 26.119 termos distintos em toda coleção. Significa que no cômputo geral houve um incremento de 2.699 termos, embora tenha havido uma redução de 218 termos distintos com a eliminação dos documentos inválidos e dos caracteres estranhos que mascaravam alguns termos. Por exemplo, o termo que estava sendo indicado erroneamente como “a cobrança”, agora é indicado corretamente como dois termos “a” e “cobrança”.

O próximo passo foi a verificação de pontuação que em muitos casos dificultam a identificação de termos, como a seguir.

Tabela 27 – Exemplo de Pontuação com Erro

Exemplo	Termo
... cliente foi debitado R$ 110,00.Diante do fato pede averiguação...	110,00.diante
... um EMPRÉSTIMO na agência 1101.Reclama, pois já se...	1101.reclama

Na tabela 27 observa-se que os termos 110,00 e diante foram unidos pelo ponto e, então, classificado como um termo. Para resolução deste problema, procedeu-se a inserção de um espaço após o sinal de ponto para que o algoritmo pudesse identificar os termos corretos.

Nesta tarefa, deve-se tomar o cuidado com alguns casos especiais como por exemplo, www.xxx.com.br que devem ser identificados como um só termo. Ainda, considerando o propósito da pesquisa, que é a criação de agrupamentos, não devem alterar o resultado final os tratamentos pontuais como o exemplo de endereço da Internet citado.

Após a retirada dos caracteres de pontuação, realizou-se a separação dos termos para verificar a quantidade e constatou-se 2.419.077 termos, o que mostra um aumento de 2.792 termos, entretanto 24.180 distintos em toda coleção. Isto mostra uma redução de 1.939 termos somente com o ajuste da pontuação.

A quantidade final de documentos da coleção que foi destinada à mineração de textos é de 51.694. Faz-se necessária a utilização de recursos mais sofisticados de lingüística computacional para a depuração dos textos e posterior aplicação ao minerador de texto.

6.6.1.6 Lematização

A maior parte do trabalho na preparação dos dados objetiva a redução do escopo do corpus, isto é, a diminuição da quantidade de termos que alimentam o minerador de textos. Um dos recursos lingüísticos disponíveis na ferramenta SAS é a redução das palavras ao Lema.

Foram lematizados 13.396 termos. Observou-se que 50% dos termos lematizados possuem apenas 2 formas, exemplo, “o” e “os” foram substituídos por “o”. Ainda que 90% dos casos, os termos possuem até 8 variantes. A tabela 28 mostra o resumo da lematização.

Tabela 28 – Estatísticas dos Termos Lematizados

N	Média	Desvio Padrão	50%	90%	Máximo
3.472	3,86	3,98	2	8	49

De acordo com os números apresentados, observa-se que a quantidade de lemas é de 3.472, o que significa uma redução de 74%, 9.924 termos, o que indica um vocabulário simples. Isto é esperado já que as mensagens são transcritas por operadores que possuem um linguajar padrão, como deseja a administração.

O número máximo de termos encontrados foi 49, esse número parece indicar um erro. Entretanto, verificando o arquivo mais detidamente, conforme a tabela 29, trata-se de um verbo comum neste tipo de texto.

Tabela 29 – Variantes de um Termo

Obs	Termo
1	informando-os
2	informar-lhe
3	informar-me
4	Informara
5	informaram-na
6	informaram-no
7	Informas
8	informassem-me
9	informo-a
10	informo-o
11	informou-me
12	informou-os
13	informou-se
14	informá-los
15	informa-o
16	Informando-a
17	Informando-o
18	Informaram-lhe
19	Informarei
20	Informassem
21	Informei
22	informa-lhe
23	Informem
24	**Informes**
25	Informou-a
26	Informou-o
27	informa-lo
28	Informarão
29	Informo
30	informou-lhe
31	Informavam
32	informa-la
33	Informadas
34	Informarem
35	Informasse
36	informá-lo
37	informá-la
38	informando-lhe
39	Informava
40	Informados
41	Informe
42	Informam
43	Informaram
44	Informar
45	Informando
46	Informou
47	Informada
48	Informado
49	Informa

Todas essas variantes serão então computadas como “informar”, logo a palavra passa a receber um peso maior segundo a sua freqüência. Observa-se que a palavra “informes” pode tanto se referir à 2ª pessoa do singular do subjuntivo (que tu) informes quanto ao substantivo plural “informes”. Essa ocorrência foi observada mais detidamente e constatou-se que se tratava do substantivo.

Nesse passo, verificou-se também a presença de muitos termos com algarismos que podem ser números de telefones, matrícula de empregado, erros de digitação, valores financeiros, número de residência, de documentos, de CEP, horários e assim por diante. Após análise, concluiu-se que esses termos não trariam informações suficientes para mantê-los na base e, então, foram descartados 1.610 termos.

6.6.1.7 Termos Freqüentes

O algoritmo de mineração de texto possui um lematizador padrão e remove todos os termos que ocorrem somente em um único documento. Nesse passo, apesar de consumir bastante tempo, a criação de listas de *stopwords* e de sinônimos são tarefas essenciais para a redução de dimensão.

Procede-se então a criação de lista de *stopwords*. Termos que se repetem em todos os documentos são pouco informativos para o objetivo de criar agrupamentos, pois não discriminam um documento de outro. Termos com peso baixo também são candidatos à eliminação. Dessa forma pesos menores que 0,10 foram eliminados. Deve-se proceder à eliminação de palavras que não foram eliminadas automaticamente, como por exemplo, nomes de pessoas, pois não se objetiva identificar ninguém e, portanto, esses termos somente aumentariam o tempo de processamento e não agregariam valor ao trabalho. Outro ponto foi unificar palavras escritas de forma errada, por exemplo, “urgente” e “urgênte". Alguns destes erros não foram identificados automaticamente e o trabalho teve que ser manual.

O programa cria alguns verbos estranhos por conta de seu algoritmo de lematização, como o caso do suposto verbo bolsar (bolsa, bolsas) e outro exemplo de um verbo que existe, Saldar, mas que foi agregado na forma de verbo indevidamente, pois os termos são os substantivos saldo e saldos.

Um outro passo para exploração dos termos com pouca capacidade de discriminação é a elaboração de agrupamentos e a verificação de suas palavras relevantes. A partir dessa base pré-limpa, foram gerados 16 agrupamentos que auxiliam a visualização dos termos relevantes, como no exemplo abaixo:

Tabela 30 – Agrupamentos para Depuração da Base

Agrupamento 2	Agrupamento 4	Agrupamento 5	Agrupamento 6
+ agência	+ alegar	+ abertura	+ agência
+ agendar	+ ar	+ abrir	+ atendimento
+ aguardar	+ atendente	+ agência	+ auto
+ atender	+ atendimento	+ cartão	+ auto-atendimento
+ atendimento	+ cidade	+ contar	+ debitar
+ concordar	+ dinheiro	+ corrente	+ depositar
+ demorar	+ efetuar	+ creditar	+ depósito
+ esperar	+ ficar	+ datar	+ dinheiro
+ ficar	+ fila	+ depositar	+ efetuar
+ fila	+ funcionar	+ enviar	+ funcionar
+ funcionar	+ funcionário	+ estornar	+ localizar
+ haver	+ jogar	+ possuir	+ máquina
+ horário	+ loteria	+ poupança	+ retirar
+ marcar	+ lotérico	+ residência	+ sacar
+ pessoa	+ pagamento	+ sacar	+ sala
+ reclamar	+ pagar	+ solicitar	+ supermercado
+ senha	+ passar	+ transferência	+ terminal
agendamento	+ providenciar	+ valor	+ utilizar
Hs	+ sacar	+ verificar	+ valor

Os termos da tabela 30 são os termos que descrevem os agrupamentos. Isto significa que esses termos são mais prováveis de ocorrer em um agrupamento e não necessariamente que ocorram em todos os documentos dentro de um agrupamento. Dessa forma, é desejável que determinado termo ocorra somente em um agrupamento, a fim de que eles sejam distintos entre si.

Observam-se aqui alguns termos que podem ser adicionados às listas de *stopwords* a fim de deixar os agrupamentos menos heterogêneos. Esse processo é repetido até que os grupos fiquem homogêneos.

Durante essas iterações, observam-se também inconsistências nas palavras compostas como, no exemplo anterior, a palavra auto-atendimento que quando grafada sem o hífen é interpretada erroneamente como auto e atendimento. Dessa forma, foi executado um trabalho de identificação de possíveis problemas dessa natureza. Inseriu-se o hífen nas palavras que deveriam ter sido grafadas com ele e o sinal de sublinhado (_) nas palavras compostas, sem hífen, para que o programa fizesse a separação dos termos de forma correta. Assim, palavra como “conta corrente” ficou “conta_corrente” e “pró jovem”, “pró-jovem”.

6.6.1.8 Amostragem

Apesar de não estar destacada na figura 12, a seleção de amostra é sempre desejável quando se manipulam grandes quantidades de registros. Nesse caso, a limitação de desempenho do equipamento utilizado, isto é, utilização intensa da máquina para processar o volume de textos, pode ser superada com o uso de amostra. Dessa forma foi extraída amostra de 7895 registros. O critério escolhido foi a amostragem estratificada pelas variáveis ‘Código do Produto’ e ‘Código do Motivo’ com o objetivo de manter as amostras com proporções semelhantes à população.

6.6.2 A Mineração de Textos

A mineração de textos abrange a criação de agrupamentos e a elaboração de modelo categorizador das novas mensagens.

6.6.2.1 Classificação

A tarefa de criação dos grupos de documentos está representada na figura 12 que indica os passos que seguidos:

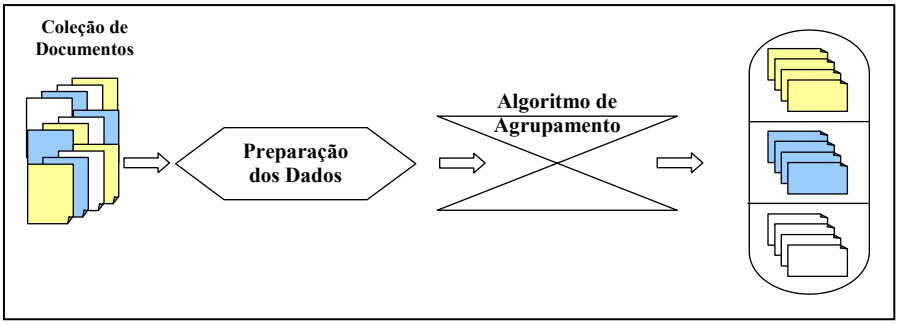

Figura 12 – Agrupamento de Documentos

O passo da preparação dos dados foi descrito na seção anterior e consiste na adequação da informação textual para o formato requerido pelo algoritmo de agrupamento. Essa etapa ocupa-se em rotular cada um dos documentos do corpus baseado no exame de seus termos e, dessa maneira, reuni-los em grupos menores que deverão conter documentos similares. A idéia é maximizar tanto a similaridade entre os documentos dentro do grupo, quanto a diferença entre eles de grupo para grupo.

A criação de grupos é um processo matemático que calcula a distância entre os documentos devidamente transformados em vetores numéricos. Portanto, os grupos criados carecem de significado, isto é, deve-se analisar cada um dos grupos para verificar o significado implícito dos documentos que contém e, assim, atribuir rótulos aos grupos para organização e identificação do assunto de cada um deles. Essa fase conta com o apoio do analista de domínio para identificação do tema de cada grupo.

Um dos objetivos específicos do projeto é verificar a aderência do processo de criação de agrupamentos automática e manual que já vem sendo executada por empregados que tratam as informações recebidas. Por meio dessa comparação foi proposta uma classificação e, a partir dela, a geração do indicador.

É importante relacionar o tipo de melhoramento que o processo de automação pode trazer, ou seja, uma das reclamações mencionadas pelos analistas

foi a ambigüidade de grupos que foram criados e nesta análise chegou-se à conclusão que dois ou mais grupos podiam ser unificados diante da constatação da similaridade entre seus documentos.

6.6.2.2 Categorização

Uma vez criados os agrupamentos automaticamente, uma tarefa cumprida foi a categorização de novos documentos que são recebidos diariamente.

Todo documento alocado em determinado grupo possui um rótulo que o identifica como membro portador das características que são específicas daquele grupo e não de outro. Esses documentos possuem padrões que os identificam e os diferenciam dos demais. A tarefa então é construir um categorizador automático baseado no conteúdo dos agrupamentos identificados.

A figura 13 ilustra o fluxo que constrói o categorizador automático, rotulado como modelagem no quadro cinza, e a entrada de novos documentos, no quadro verde identificado como categorização, que passam pelo categorizador que os distribui entre os grupos segundo a similaridade de seu conteúdo.

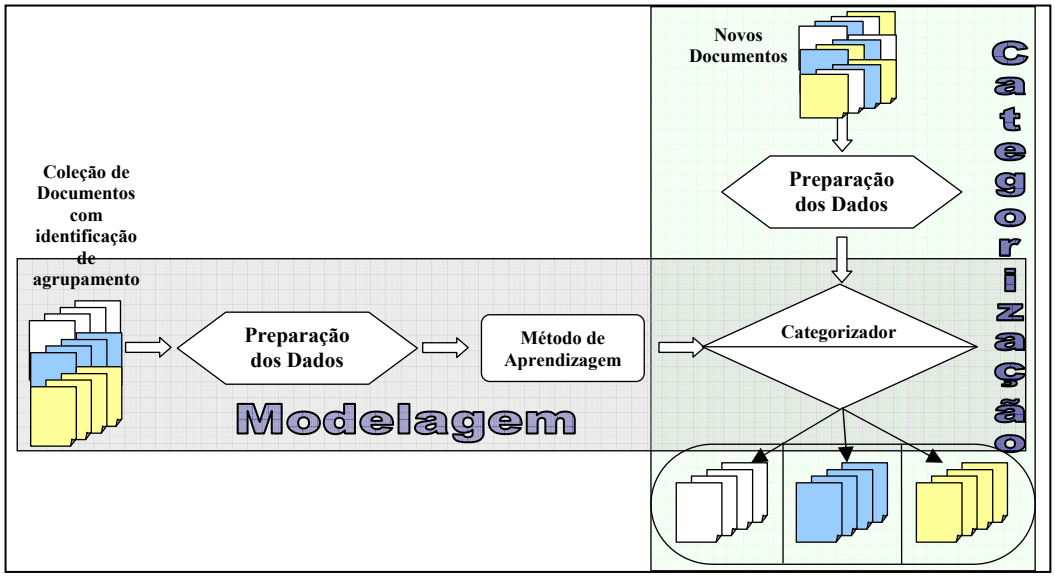

Figura 13 – Categorização de documentos

A modelagem é baseada em processos já bem estabelecidos principalmente na Estatística e na Inteligência Artificial. Nestes processos, espera-se que os dados passados e acumulados possam indicar alguns critérios de decisão para

categorização, isto é, que se aprenda, a partir dos dados, a estrutura implícita que caracteriza a alocação de um documento em um determinado grupo.

O algoritmo de categorização, chamado de categorizador, tem como entrada os novos documentos que passam pelos processos de transformação de texto em número. Assim, pela análise de peculiaridades do novo documento e comparação com sua aprendizagem na coleção para distinção de grupos, o categorizador indicará sua provável categorização.

6.7 Criação do Indicador

Considerando que uma base SAC contempla a insatisfação dos clientes, é razoável supor que produtos ou serviços com grande número de queixas estejam com problemas e devem ser analisados a fim de resolver o desagrado por parte da clientela.

Os grupos foram formados por similaridade entre os documentos e, então, contemplaram, de maneira organizada e estruturada, os principais tópicos. A convergência de reclamações para determinados grupos destaca possíveis problemas pontuais em locais específicos, lançamento de produtos mal planejados, serviços deficientes e problemas de pessoal.

O indicador foi formulado através de índices que destacam agrupamentos de produtos ou serviços de acordo com suas ocorrências. Esses índices podem servir de indicativos para uma verificação mais detalhada de pontos críticos.

Seja G o agrupamento onde

G_1 = Atendimento	G_5 = Internet	G_9 = Cadastro
G_2 = agencia	G_6 = canais	G_{10} = Empréstimo
G_3 = cartão, G1	G_7 = outros	G_{11} = FGTS
G_4 = habitação	G_8 = P. Social	G_{12} = Cheque

O indicador para o acompanhamento é:

Equação 7

$$IS_i = \frac{G_i}{\sum_i G_i} x100$$ **onde**

IS_i representa o índice de satisfação no agrupamento *i*

G_i é a soma de todas as ocorrências de reclamação no agrupamento *i*

$\sum_i G_i$ é a soma total das ocorrências de reclamação.

Por exemplo, se G_1 é igual a 260 ocorrências em um total de 1000 ocorrências, tem-se $IS_1 = \frac{260}{1000} x100 = 26$, isto é, 26% para o agrupamento G_1.

Os pontos críticos poderão ser visualizados através de gráficos e tabelas que fornecem subsídios importantes para os gestores que prontamente podem tomar decisões na resolução do problema, como adequação de serviço ou produto, treinamento de pessoal, substituição de pessoas e assim por diante.

7 RESULTADOS

De posse de dados e ferramentas apropriadas para o desenvolvimento do projeto, foram realizadas as tarefas descritas na metodologia e que se apresentam no decorrer deste capítulo.

7.1 Ponderação dos Termos

Dado que o trabalho de depuração da base se encontrava em um nível satisfatório, realizou-se o agrupamento da coleção com o objetivo de alocar os documentos semelhantes em grupos. O objetivo é maximizar a diferença entre os grupos e minimizar a diferença internamente.

A criação desses grupos é dependente da escolha do critério de transformação de textos em números, isto é, a ponderação de termos. Aproveitando a facilidade do programa, utilizaram-se 4 versões de ponderações disponíveis, Entropia, GF/IDF, IDF, Normal (ver tabela 10) e, também, a não aplicação de nenhuma ponderação, isto é, somente a freqüência simples dos termos.

Observando-se a variação interna, destacam-se os melhores resultados que foram: Entropia e DF. A transformação utilizando a Entropia fornece 18 agrupamentos, enquanto a IDF fornece 15.

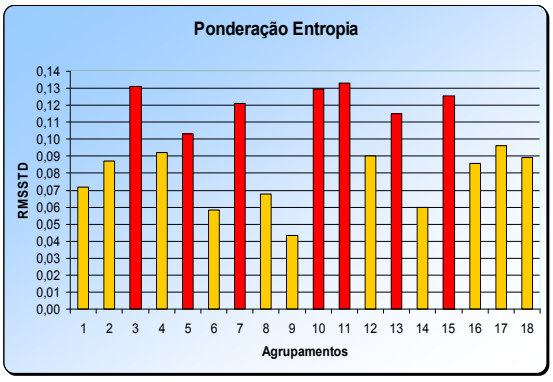

Figura 14 – Entropia

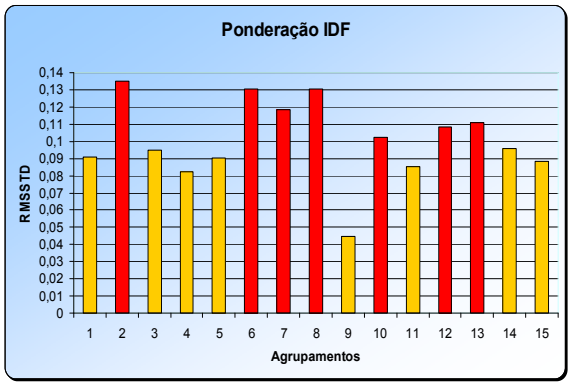

Figura 15 – IDF

A quantidade de grupos que estão com variabilidade menor que 10% na Entropia é de 61% (11 em 18), enquanto na IDF é de 53% (8 em 15), além disso existem mais agrupamentos na Entropia e uma maior quantidade, 5, de grupos abaixo de 8%. Isto indica a construção de grupos mais homogêneos e, portanto, a

Entropia foi o tipo de ponderação escolhido para a identificação dos agrupamentos e comparação com a classificação manual.

7.2 Classificação Automática X Manual

A classificação manual foi realizada utilizando a combinação de “Assunto”, “Produto e Serviço” e “Motivo” que totalizam 1517 agrupamentos. Entretanto, considerando somente as reclamações, tem-se 785 agrupamentos. A classificação é realizada com o critério subjetivo do operador que “decide” em qual destas classes a mensagem será enquadrada. O primeiro nível, **Assunto**, possui 19 grupos, embora existam 3 grupos sem descrição que foram alocados em um único agrupamento denominado Outros, e é a primeira classificação que o operador faz. Em seguida, classifica-se a mensagem no segundo nível e, então no terceiro para, assim, enviá-la para a área responsável pelo fornecimento de respostas, quando for o caso.

Nesse estudo, realiza-se a classificação automática do primeiro nível e compara-se com a classificação manual para aferir a qualidade da criação de agrupamentos automaticamente com base no conteúdo das mensagens.

Tabela 31 – Comparação Classificação Manual X Automática

Grupo Assunto X Agrupamento																				
Código do Grupo Assunto		Agrupamento (Cluster ID)																		Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
Loterias	Qtde	9	0	72	22	14	0	1	0	3	12	39	0	6	0	69	0	2	4	253
	% linha	4	0	28	9	6	0	0	0	1	5	15	0	2	0	27	0	1	2	
	% coluna	3	0	7	4	4	0	0	0	2	1	4	0	2	0	7	0	1	2	
Fgts	Qtde	0	0	11	0	0	0	0	0	0	56	27	2	2	0	7	1	0	1	107
	% linha	0	0	10	0	0	0	0	0	0	52	25	2	2	0	7	1	0	1	
	% coluna	0	0	1	0	0	0	0	0	0	6	3	1	1	0	1	0	0	0	
Serviços Bancários e arrecadação	Qtde	0	0	33	1	2	0	1	0	0	10	24	0	9	0	8	3	6	10	107
	% linha	0	0	31	1	2	0	1	0	0	9	22	0	8	0	7	3	6	9	
	% coluna	0	0	3	0	1	0	0	0	0	1	2	0	3	0	1	1	2	4	
Marketing	Qtde	0	0	1	0	0	0	0	0	0	2	1	0	0	0	2	1	0	0	7
	% linha	0	0	14	0	0	0	0	0	0	29	14	0	0	0	29	14	0	0	
	% coluna	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Agência	Qtde	65	41	127	388	35	10	40	7	6	184	266	6	50	11	568	37	53	73	1967
	% linha	3	2	6	20	2	1	2	0	0	9	14	0	3	1	29	2	3	4	
	% coluna	21	16	12	72	10	6	9	4	4	20	27	3	15	4	55	14	21	32	
Internet	Qtde	1	5	64	0	0	0	0	0	0	2	3	0	5	0	8	1	0	0	89
	% linha	1	6	72	0	0	0	0	0	0	2	3	0	6	0	9	1	0	0	
	% coluna	0	2	6	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	
Disque caixa	Qtde	113	102	119	105	21	126	56	0	7	54	170	4	67	0	102	1	25	20	1092
	% linha	10	9	11	10	2	12	5	0	1	5	16	0	6	0	9	0	2	2	
	% coluna	36	40	11	20	6	79	13	0	5	6	17	2	21	0	10	0	10	9	
Produtos de fidelização	Qtde	0	0	11	0	0	0	0	0	0	56	27	2	2	0	7	1	0	1	107
	% linha	0	0	10	0	0	0	0	0	0	52	25	2	2	0	7	1	0	1	
	% coluna	0	0	1	0	0	0	0	0	0	6	3	1	1	0	1	0	0	0	
Recursos humanos	Qtde	1	0	4	0	0	0	0	0	0	1	3	0	3	0	5	1	0	0	18
	% linha	6	0	22	0	0	0	0	0	0	6	17	0	17	0	28	6	0	0	
	% coluna	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
Canais de atendimento	Qtde	5	0	48	10	207	2	4	0	0	8	21	2	2	0	27	0	3	20	359
	% linha	1	0	13	3	58	1	1	0	0	2	6	1	1	0	8	0	1	6	
	% coluna	2	0	4	2	61	1	1	0	0	1	2	1	1	0	3	0	1	9	
Outros	Qtde	0	0	1	0	0	0	0	0	0	1	1	0	2	0	4	0	1	0	10
	% linha	0	0	10	0	0	0	0	0	0	10	10	0	20	0	40	0	10	0	
	% coluna	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
Empréstimo	Qtde	1	0	15	1	1	2	3	0	0	36	27	1	5	0	14	188	4	5	303
	% linha	0	0	5	0	0	1	1	0	0	12	9	0	2	0	5	62	1	2	
	% coluna	0	0	1	0	0	1	1	0	0	4	3	1	2	0	1	71	2	2	
Cartões	Qtde	36	15	194	1	22	5	191	0	5	86	65	156	26	250	31	1	0	3	1087
	% linha	3	1	18	0	2	0	18	0	0	8	6	14	2	23	3	0	0	0	
	% coluna	11	6	18	0	6	3	44	0	3	9	7	82	8	89	3	0	0	1	
Habitação	Qtde	23	0	51	4	0	2	77	152	0	315	71	14	30	3	35	14	4	2	797
	% linha	3	0	6	1	0	0	10	19	0	40	9	2	4	0	4	2	1	0	
	% coluna	7	0	5	1	0	1	18	95	0	34	7	7	9	1	3	5	2	1	
Captação/ contas	Qtde	39	0	151	0	35	3	30	1	2	84	135	5	42	17	48	15	9	92	708
	% linha	6	0	21	0	5	0	4	0	0	12	19	1	6	2	7	2	1	13	
	% coluna	12	0	14	0	10	2	7	1	1	9	14	3	13	6	5	6	4	40	
Transf. De benef./ prog. Sociais	Qtde	3	61	34	1	4	1	14	0	124	39	35	0	44	1	19	1	8	0	389
	% linha	1	16	9	0	1	0	4	0	32	10	9	0	11	0	5	0	2	0	
	% coluna	1	24	3	0	1	1	3	0	84	4	4	0	14	0	2	0	3	0	
Total		317	253	1069	536	342	159	435	160	148	917	972	190	324	282	1040	264	256	231	7895

A tabela 31 mostra na primeira coluna as classes criadas pelos operadores e as colunas de 1 a 18 são as criadas pelo algoritmo de classificação. Cada célula

representa a intersecção de uma linha com uma coluna. Por exemplo, a linha "Loterias" Qtde com a coluna 1 possui 9 registros em comum; com a coluna 2, nenhum registro em comum, isto é, 0; e assim sucessivamente, até cobrir todas as colunas. A última coluna apresenta a soma da linha. As linhas com a descrição "% linha" representam o percentual da classificação manual em relação à classificação automática. Na primeira coluna o número 4 significa que das mensagens classificadas manualmente como "Loterias", em um total de 253, apenas 4% ou 9 de 253 foram classificadas automaticamente na coluna 1. As linhas com a descrição "% coluna" representam o percentual da classificação automática em relação à manual. Por exemplo, na coluna 1 com a terceira linha o número 3 significa que dos 317 registros classificados automaticamente na coluna 1 apenas 3% também foram classificados manualmente em Loterias.

Os agrupamentos 1, 2, 4, 6, 8, 9, 12, 14, 16, 17 e 18 (em vermelho) são os grupos com menor variabilidade interna. Isto indica que os documentos destes grupos devem ser homogêneos. Os quadros contendo os percentuais, de cor laranja, indicam onde houve maior coincidência entre a classificação automática e a classificação manual.

A primeira observação é que a classificação manual possui 16 categorias, enquanto a automática 18, portanto, não há a perfeita concordância entre as duas classificações.

O agrupamento 8 foi classificado 95% das vezes na categoria "Habitação" e, conseqüentemente, esse grupo deve tratar de assuntos relacionados à habitação. As palavras descritoras deste grupo são: + hipotecar, + cartório, + documento, + liquidar, imobiliário.

O sinal de "+" indica que o termo foi lematizado. Esses termos parecem, de fato, ser relacionados ao tema habitação e poderiam identificar o agrupamento com o título Habitação. Contudo, observa-se no quadro que o termo habitação escolhido pelos operadores é mais amplo, pois apenas 19% dos documentos foram classificados no grupo 8. Alem disso, o grupo 10 também contempla 34% de documentos coincidentes com a categoria manual Habitação, sendo que 40% da classificação dos operadores que classificaram os documentos como Habitação também coincide com o grupo 10. Dessa forma, do ponto de vista da classificação

manual, os agrupamentos 8 e 10 tratam de assuntos específicos distintos do tema habitação.

Tabela 32 – Exemplo de Agrupamentos

agrupamento 8	agrupamento 10
a cliente relata que quitou o seu contrato de gaveta em nome de XXXXX XXXXXX em 18 / 12 / 2000, e até o dado momento não recebeu o termo de quitação para liberação da hipoteca junto ao cartório. não compreende o porque da demora de entrega de seu documento. acrescenta que já procurou a agência várias vezes e ninguém soluciona este problema. pede providências com urgência.	a cliente relata que tem um contrato pelos sistema pa . salienta ainda que o condomínio é administrado link parque, que não faz um bom trabalho e nenhuma melhoria. não concorda com o aumento de 60% e solicita providências urgentes.
a cliente relata que quitou um financiamento imobiliário número XXXXXXXXX em 19. 12. 05. acrescenta que gostaria de receber a baixa da hipoteca. pede providências.	cliente relata possui um contrato habitacional, o qual deste 19 / 09 / 2004 está tentando liquidar o imóvel com fgts, ressalta foi deixado todos os documentos para liquidação, porém até o presente momento não houve êxito. diante dos fatos pede providências com urgência.

De acordo com os documentos na base exemplificados na tabela 32, infere-se que o agrupamento 8 trata mais especificamente de assuntos habitacionais relacionados a liberação de hipoteca e o agrupamento 10 com assuntos mais diversificados do termo habitação. Essa especificidade do grupo 8 e maior generalização do grupo 10 pode ser percebida na variação interna apresentada anteriormente no gráfico 15.

Um outro exemplo é o agrupamento 16 que possui as palavras descritoras: + descontar, + aposentar, + folha, + consignar, + empréstimo. Pelas palavras descritoras tem-se um indicativo que se trata de assunto relacionado a empréstimo. Comparando com a classificação manual temos 71% de coincidência da classificação automática em relação à manual e 62% da manual para automática. Dessa forma, pode-se seguramente afirmar que o tema é Empréstimo e a confirmação pode ser feita na consulta à base de documentos que realmente mostra que a inferência está correta.

Um exemplo de classificação que não parece coincidir com nenhuma classificação manual é o agrupamento 1, pois não se observa uma concentração em um assunto específico. As palavras descritoras do grupo são: + telefonar, + desligar, + cara, atendimento, + nervoso.

Esse grupo ficou distribuído majoritariamente em 21% (Agência), 36% (Disque-Caixa), 11% (Cartões) e 12% (Captação/Contas). Olhando-se as palavras descritoras e comparando-se a distribuição com as classificações manuais parece ser contraditório, pois as palavras indicam um tema relacionado ao atendimento por meio do telefone.

Esse fato pode servir de indicador para treinamento do quadro de operadores visando padronizar a recepção das mensagens para facilitar a tarefa de identificação do tema e de remessa à área gestora de determinado assunto para o seu encaminhamento.

O processo de identificação do tema dos agrupamentos é realizado da mesma maneira para todos os restantes, isto é, tenta-se identificar o assunto pelas palavras descritoras e, caso necessário, verifica-se o conteúdo de mensagens alocadas para o agrupamento em questão visando assegurar o entendimento do tema que os textos abordam.

A classificação final é a seguinte:

Tabela 33 – Identificação dos Agrupamentos

grupo 1	grupo 2	grupo 3	grupo 4	grupo 5	grupo 6	grupo 7	grupo 8	grupo 9
+ telefonar	pis	+ página	+ lento	+ fechar	+ insistente	+ condomínio	+ hipotecar	bolsa_família
+ desligar	incorreto	+ resolver	+ insuficiente	auto-atendimento	+ dificuldade	Administração	+ cartório	+ benefício
Cara	+ rendimento	+ site	+ preferencial	+ sala	+ orientação	+ bloquear	+ contrato	+ estudar
+ atendimento	+ consultar	+ técnico	+ agendar	+ supermercado	Atendente	+ endereço	+ financiar	pró-jovem
+ nervoso	+ solução	+ Internet	ruim	+ equipamento	Telefonista	+ gravação	+ imobiliário	cadastramento
Atendimento Telefone	**Serviço Agência**	**Tecnologia Internet**	**Atendimento Agência**	**Canais de Atendimento**	**Disque-Caixa**	**Outros**	**Habitação**	**Programas Sociais**

grupo 10	grupo 11	grupo 12	grupo 13	Grupo 14	grupo 15	grupo 16	grupo 17	grupo 18
+ contrato	+ travar	+ vencimento	+ alteração	abertura	+ educar	+ empréstimo	+ rescisão	+ talão
+ financiar	porta_giratória	+ cancelamento	cep	+ universitário	+ grosseiro	+ aposentar	Cnpj	+ compensar
+ vender	+ procedimento	cartão_de_crédito	+ correspondência	+ viajar	+ tratar	+ folha	+ demitir	+ devolver
+ construção	+ esperar	+ cancelar	+ atualizar	+ idade	+ descortês	+ desconto	+ empresa	+ microfilmagem
+ aprovar	+ segurança	anuidade	incorreto	+ poupança	+ profissional	+ consignação	+ recolher	+ tarifa
Financiamento Construção	**Acesso Agência**	**Cartão de Crédito**	**Cadastro**	**Cartão de Débito**	**Atendimento Empregado**	**Empréstimo**	**FGTS**	**Cheque**

A análise destes grupos mostra que alguns temas foram divididos em mais de um grupo, como Atendimento nos grupos 1, 4 e 15 , o que na classificação manual foi agrupado em apenas um. Observando-se os textos destes grupos constata-se que realmente se encaixam no tema Atendimento, porém com as nuances que os diferencia conforme a identificação dos agrupamentos. Por exemplo, o grupo 1 trata especificamente de atendimento por intermédio do telefone, enquanto que o grupo 4, do atendimento em geral na agência e o grupo 15 aborda mais especificamente o atendimento do empregado.

Tabela 34 – Agrupamentos Automáticos Finais

Atendimento	grupo 1	+ telefonar	+ desligar	cara	+ atendimento	+ nervoso	4%
	grupo 4	+ lento	+ insuficiente	+ preferencial	+ agendar	Ruim	7%
	grupo 15	+ educar	+ grosseiro	+ tratar	+ descortês	+ profissional	13%
	grupo 6	+ insistente	+ dificuldade	+ orientação	atendente	Telefonista	2%
Agência	grupo 2	Pis	incorreto	+ rendimento	+ consultar	+ solução	3%
	grupo 11	+ travar	porta_giratória	+ procedimento	+ esperar	+ segurança	12%
Cartão	grupo 12	+ vencimento	+ cancelamento	cartão_de_crédito	+ cancelar	Anuidade	2%
	grupo 14	Abertura	+ universitário	+ viajar	+ idade	+ poupança	4%
Habitação	grupo 8	+ hipotecar	+ cartório	+ contrato	+ financiar	+ imobiliário	2%
	grupo 10	+ contrato	+ financiar	+ vender	+ construção	+ aprovar	12%
Internet	grupo 3	+ página	+ resolver	+ site	+ técnico	+ Internet	14%
Canais	grupo 5	+ fechar	auto-atendimento	+ sala	+ supermercado	+ equipamento	4%
Outros	grupo 7	+ condomínio	administração	+ bloquear	+ endereço	+ gravação	6%
P.Social	grupo 9	bolsa_família	+ benefício	+ estudar	pró-jovem	Cadastramento	2%
Cadastro	grupo 13	+ alteração	cep	+ correspondência	+ atualizar	Incorreto	4%
Empréstimo	grupo 16	+ empréstimo	+ aposentar	+ folha	+ desconto	+ consignação	3%
FGTS	grupo 17	+ rescisão	cnpj	+ demitir	+ empresa	+ recolher	3%
Cheque	grupo 18	+ talão	+ compensar	+ devolver	+ microfilmagem	+ tarifa	3%

A tabela 34 contempla os agrupamentos consolidados por tema. Com essa junção, obteve-se 12 grupos. Do ponto de vista prático, a junção destes agrupamentos correlatos em um único tema é desejável.

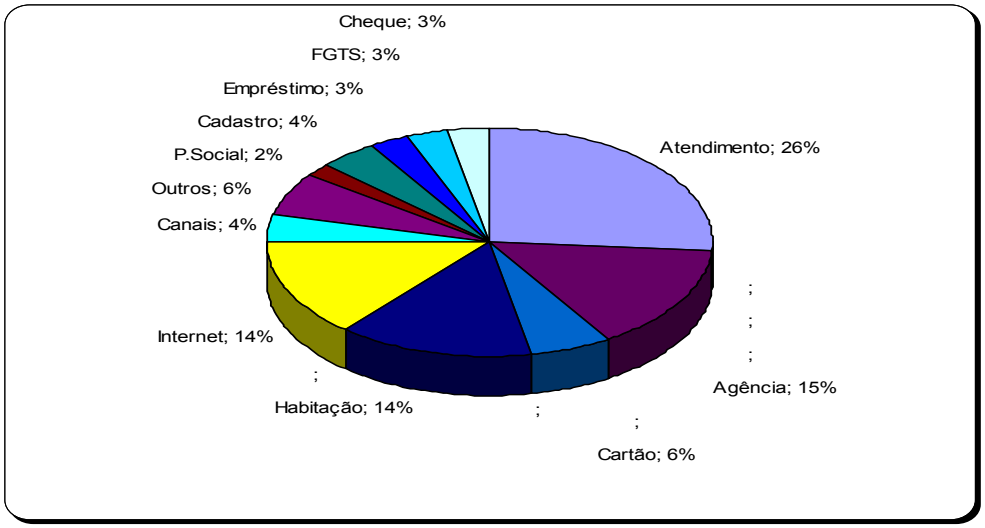

Figura 16 – Representação da Proporção dos Agrupamentos

A figura 16 apresenta os agrupamentos finais com suas respectivas proporções.

7.3 Indicador

O objetivo da estruturação da informação textual contida no SAC culmina com a criação de indicadores de acompanhamento. Para que se faça uma gestão eficiente da performance dos produtos e serviços frente aos consumidores é necessário à elaboração de critérios que possibilitem a mensuração da satisfação do cliente.

Aplicando a equação 7 na base temos o desempenho mensal por agrupamento.

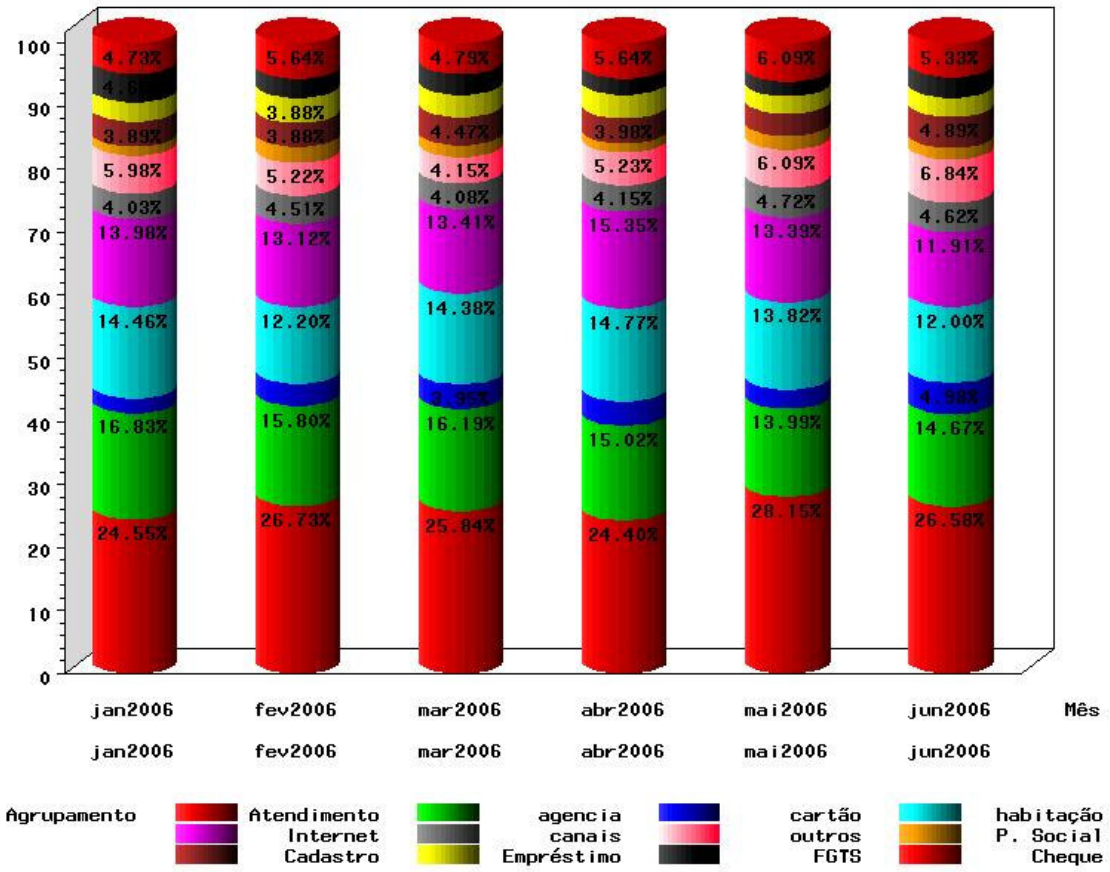

Figura 17 – Acompanhamento Mensal

No gráfico da figura 17, pode ser observado mensalmente tanto o comportamento geral, quanto o dos agrupamentos separadamente. Nos agrupamentos pode-se perceber, por exemplo, a diminuição do agrupamento Internet que em janeiro estava com 13,98% de todas as ocorrências e em junho com 11,91%, isto representa um decréscimo de 2,07%. Por outro lado, no agrupamento

Atendimento aponta uma leve tendência de crescimento das ocorrências de 24,55% em 26,58%.

Com esse tipo de indicador a Administração atenta pode reverter a insatisfação do cliente com intervenção pontual nos agrupamentos que forem se destacando no conjunto.

7.4 Categorizador Automático

Uma tarefa desejada é que as mensagens recebidas sejam automaticamente classificadas e, para tanto, é necessário o desenvolvimento de um modelo preditivo que as classifique sem a intervenção humana.

Existem várias opções de ferramentas para construção de modelos preditores no programa utilizado neste estudo. Alguns usam métodos estatísticos e outros não. O fato é que nenhuma técnica resolve todos os problemas na mineração de dados e que cada uma delas possui o seu ponto forte e fraco. A escolha do melhor modelo é dependente da aplicação e deve ser feita baseado em medidas que validam a qualidade do modelo, isto é, o quão assertivo é o resultado produzido.

Para a construção do modelo de classificação neste trabalho, foram usadas 4 opções disponíveis no programa SAS: regressão logística, árvore de decisão, *memory-based reasoning* e rede neural

Dessa forma, com base no histórico de mensagens recebidas e suas classificações espera-se que um modelo capture a associação dos termos contidos nas mensagens com os agrupamentos nos quais eles foram alocados. Então, um modelo de classificação recebe novos registros e atribui uma classificação já existente a eles.

Para o desenvolvimento do modelo, uma prática comum é dividir os dados de modo que se possa sempre validar o modelo com dados que não foram utilizados para sua construção. Assim, a amostra foi dividida em três partes: para treinamento, para validação e para teste do modelo.

A primeira, treinamento, com 40% do total de documentos foi usada para construção do modelo inicial, a segunda, teste, 30%, foi utilizada para ajustar o modelo inicial e torná-lo mais geral, isto é, evitar que o modelo seja efetivo somente

se aplicado à base de treinamento. A validação foi a terceira e última parte utilizada para medir a provável efetividade do modelo em dados novos, isto é, os 30% de dados restantes que não foram utilizados na construção dele. A divisão foi feita utilizando a estratificação pelas variáveis produto e grupo assunto.

Os resultados das taxas de erro na classificação de cada modelo estão na tabela 35 que se segue:

Tabela 35 – Taxa de Classificação dos Modelos

Name	Treinamento: Taxa De Erro Na Classificação	Teste: Taxa De Erro Na Classificação	Validação: Taxa De Erro Na Classificação
Memory	23%	25%	24%
Logística	5%	16%	14%
Arvore	49%	53%	51%
neural	5%	15%	13%

A leitura dessa tabela diz que na fase de treinamento as melhores performances foram dos modelos de regressão logística e rede neural, 5% cada um. Isto quer dizer que os modelos erram em 5% dos casos de classificação, ou seja, o valor predito pelo modelo não confere com classificação realizada na criação dos agrupamentos. Na fase de teste, na qual o modelo é ajustado para generalizar a previsão, as taxas passam para 16% e 15% para Regressão Logística e Rede Neural, respectivamente. Nota-se que a diferença entre os modelos é muito pequena. Na última fase, aplicação em uma base de dados novos para validação o resultado é 14% para Regressão Logística e 13% para Rede Neural.

Considerando os resultados apresentados, o modelo de Rede Neural apresentou a melhor performance e, portanto, é o candidato a ser implementado para a classificação de novos registros que serão recebidos na base.

Vale ressaltar que todo modelo preditivo deve passar por uma reavaliação após um determinado período de uso, pois o seu poder de previsão pode deteriorar drasticamente, caso ocorram mudanças significativas nos dados.

A utilização da classificação automática de textos pode trazer mais velocidade ao tratamento das novas mensagens recebidas diariamente visto que as tarefas de recepção e classificação teriam menor intervenção humana do que a do modelo operacional adotado atualmente. O erro resultante da classificação automática sugere que uma parcela de empregados, hoje utilizados nos mutirões para reclassificação das mensagens, execute a tarefa de depuração destes

agrupamentos e assinalem as inconsistências encontradas para que possam realimentar o modelo tornando sua depuração contínua e visando a redução do erro a cada iteração.

De forma geral, o modelo demonstra sua utilidade na sua implementação plena, isto é, da recepção da mensagem até o seu envio à área responsável, por não depender de treinamento de atendentes para realizar sua classificação e o treinamento e implementação do modelo ficar centralizado e controlado pelo gestor e por uma pequena equipe de desenvolvedores. Assim, as alterações que visam o melhoramento do modelo automático independem da curva de aprendizado dos atendentes. Suas tarefas, hoje restritas à leitura e classificação dos textos, migrariam para atividades que atendam as necessidades da administração.

8 CONCLUSÃO

O trabalho desenvolvido aborda o tema de tratamento automático da informação textual que objetiva a menor intervenção humana possível. Almeja-se com isso a liberação de recursos humanos para atividades intelectuais que a máquina ainda não foi preparada para fazer.

Percebe-se que o desenvolvimento tecnológico auxilia na velocidade e no volume de tratamento de dados. Porém, a informação textual ainda carece de profissionais e ferramentas, utilizadas em larga escala, capazes de manuseá-la com a mesma destreza das informações em formato de bancos de dados ou, comumente, chamadas informações estruturadas.

Para atingir os objetivos propostos utilizou-se a metodologia da DCT que vai desde a escolha da base de dados até a utilização efetiva da informação descoberta que se transforma em conhecimento diante das interpretações humanas para aplicação de forma prática.

A base de dados utilizada não estava pronta para mineração e apresentou vários problemas. Foi preciso a construção de programa para importação desses dados, no qual o suporte SAS foi de grande valia. Já com os dados importados, um outro problema, a duplicação de registros, demandou muito tempo para eliminá-los.

Para tratamento dos textos, os erros de ortografia e os erros de pontuação comprometem o resultado final e, portanto, sua resolução constitui-se num ponto crítico do trabalho, no qual a intervenção humana interativa e iterativa é uma necessidade.

Durante a etapa de descrição da base, foi possível visualizar o potencial de negócio que a informação das necessidades do cliente, na forma de reclamação, suscita. Uma vertente é o canal direto do cliente com a Empresa que pode agir pontualmente baseada nas observações escritas e melhorar seus processos internos e oferecer melhores produtos e serviços. Se a Instituição atende a reclamação do cliente, ele se sente respeitado e estabelece uma relação duradoura que culmina na realização de muitos negócios vantajosos para ambos.

O programa SAS foi utilizado em todo o processo. A ferramenta é conhecida no meio de profissionais de estatística e apresenta um leque de ofertas de procedimentos bem variado, além de uma poderosa linguagem de programação. Um dos pontos fracos da ferramenta está na dificuldade de instalação que não é trivial, pois não fornece manuais em português e os procedimentos que devem ser realizados após a instalação normal é de um grau mediano de complexidade. São feitas alterações nos arquivos de configuração e algumas exclusões de arquivos que apesar de serem instalados automaticamente comprometem o bom funcionamento do programa.

Outro problema encontrado no programa, porém específico do tema tratado na pesquisa, que é a mineração de textos, é o de adaptação à língua portuguesa. O trabalho realizado com a ferramenta de mineração de textos, "*text miner"*, é pioneiro no Brasil e foi utilizado pela primeira vez na língua portuguesa e todo pioneirismo tem seu preço. Os caracteres especiais da língua portuguesa como, por exemplo, "ç ~ é à" que não existem no inglês, ocasionaram erros no reconhecimento dos termos. Diante disso, o fabricante lançou nova versão do produto que resolveu o problema descrito. Entretanto, isso tomou muito tempo, de outubro a janeiro, entre o lançamento da nova versão e problemas relacionados à instalação.

Verificou-se, também, que algumas opções importantes existentes para língua inglesa não estão disponíveis para o português, como por exemplo o reconhecimento automático de entidades que são pessoas, lugares e instituições conhecidas.

Os pontos fortes da ferramenta são vários e se destaca a capacidade de manipulação de grande volume de registros e a disponibilidade de procedimentos e algoritmos complexos. O SAS é dotado de uma linguagem de programação proprietária bastante flexível, por isso, o profissional com habilidade em técnicas de programação pode desenvolver um sistema adequado às suas necessidades. Para aqueles que não dominam as técnicas de programação, todo o trabalho pode ser feito por interface gráfica intuitiva que oferece um rol de procedimentos que abrange todo o processo da DCT. A facilidade de utilização algoritmos complexos, como rede neural e análise de componentes principais, por exemplo, é um dos pontos fortes do SAS.

Considera-se que o objetivo dessa pesquisa foi alcançado, pois a proposta era extrair conhecimento da base SAC, criar agrupamentos automáticos com utilização de ferramenta de mineração de texto, comparar os agrupamentos criados automaticamente e manualmente, criar modelo de classificação de automática das novas mensagens recebidas e propor indicador que reflete o grau de satisfação do cliente em relação aos produtos e serviços oferecidos.

A tabela 36 sintetiza os objetivos específicos propostos e os resultados encontrados com o desenvolvimento desse trabalho.

Tabela 36 – Comparação entre objetivos específicos X resultados

OBJETIVOS ESPECÍFICOS	***RESULTADOS***
Extrair conhecimento da base SAC com aplicação da DCT;	Estatísticas Descritivas sobre o Texto, número de termos, evolução anual e mensal dos registros, quantidade média de termos por documento, desvio-padrão, além problemas destacados na pesquisa
Identificar e/ou propor categorias de agrupamento de tipos de reclamações com base no conteúdo das reclamações e respostas;	Identificados 18 agrupamentos inicialmente que foram rearranjados em 12
Analisar se a classificação automática obtida por meio da DCT possibilita otimização da classificação humana atualmente adotada;	A comparação não encontrou correlação entre todos os grupos criados manualmente e automaticamente
Criar modelo capaz de classificar automaticamente novos documentos com indicação de assertividade;	Modelos testados utilizando memory-based reasoning, Regressão Logística, Árvore de Decisão e Rede Neural(que apresentou o melhor resultado)
Gerar indicador relativo à satisfação dos clientes para subsidiar políticas e estratégias de atendimento.	Indicador criado: $IS_i = \frac{G_i}{\sum_i G_i} x100$

A metodologia proposta foi útil e aplicável em transformar base de dados em formato textual em informações organizadas, na extração de conhecimento e na automatização de processos que dependem de leitura de pessoas dedicadas a essa tarefa.

No âmbito acadêmico, considera-se que a pesquisa obteve êxito, porém sua aplicação no âmbito profissional precisa de equipe dedicada. Tal tipo de trabalho requer uma integração diária com os vários profissionais que compõem este

segmento, isto é, do atendente, dos especialistas de domínio, da equipe de tecnologia e dos pesquisadores dedicados à elaboração de modelos e de indicadores. O sucesso desse trabalho requer investimentos em treinamento específico direcionado aos atendentes para tarefa de padronizar a entrada do texto com o objetivo de minimizar o esforço de pré-processamento, aos especialistas de domínio para criação de um vocabulário elaborado com o propósito de facilitar a identificação do tema da mensagem, aos mantenedores das bases de dados para a realização de trabalhos expressivos na limpeza e preparação delas e aos pesquisadores para capacitá-los e atualizá-los de novas tecnologias, de técnicas e de metodologias. Além disso, promover a aproximação contínua entre equipe e gestores do negócio buscando traduzir na forma de modelos e indicadores os anseios da cúpula da organização.

8.1 TRABALHOS FUTUROS

Como mencionado, o trabalho de mineração de textos em língua portuguesa está apenas no começo e há muito que ser pesquisado. A aplicação da DCT na Recuperação da Informação - RI pode possibilitar um ganho na qualidade dos motores de busca.

Uma das principais características da RI é o cálculo de similaridade entre a consulta submetida e a coleção recuperada. Em um primeiro estágio podem ser recuperados os documentos similares, apenas buscando-se palavras chaves. Essa é a maneira mais usual de pesquisa dos motores de busca. Num segundo passo, a consulta é examinada e transformada em valores que serão comparados às medidas da coleção recuperada. Quando o texto é transformado em peso e reduzido aos vetores singulares que representam o texto, uma parte do contexto semântico fica latente. Dessa forma, a utilização desses vetores para medir a similaridade entre a consulta submetida e a coleção de documentos recuperados deve contemplar o problema da semântica dos termos da pesquisa, pois a busca seria nos agrupamentos criados pelo processo da DCT na coleção recuperada.

Do ponto de vista prático e comercial, há considerações a serem respondidas. A velocidade de recuperação precisa ser factível, caso contrário inviabiliza qualquer possibilidade de oferecer o serviço aos usuários. O investimento deve ser

considerado como qualquer transação realizada no mercado. Por fim, o aparato tecnológico que deverá equipar esses motores de busca para que façam o trabalho esperado, a recuperação da informação que o usuário deseja no tempo que ele almeja.

REFERÊNCIAS

AGRAWAL, R.; PSAILA, G. 1995. **Active Data Mining**. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 3–8. Menlo Park, Calif.: American Association for Artificial Intelligence. [1]

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. ACM Press Series/Addison Wesley, New York, May 1999. [1] [2]

BARLOW, J. & MOLLER, C. **Reclamação de cliente? Não tem melhor presente**. São Paulo: Futura, 1996. [1] [2]

BERRY, M. J. A.; LINOFF, G. **Data mining techniques** - for marketing, sales, and customer support. John Wiley & Sons, New York, 1997. [1] [2] [3] [4] [5]

BRÄSCHER, M. **Tratamento automático de ambigüidades na recuperação da informação**. 290 f. Tese (Doutorado em Ciência da Informação) - Curso de Pós-graduação em Ciência da Informação, Universidade de Brasília, Brasília, 1999. [1]

BREDER, I. **O Sistema de Atendimento ao Cliente – SAC e o Papel Desempenhado na Tomada de Decisão Estratégica da Caixa**. 93 f. Monografia (Especialização em Marketing Bancário) - Curso de MBA em Marketing Bancário, Fundação Getúlio Vargas, Brasília, 2002. [1] [2]

CHAUVEL, M. A. **Consumidores insatisfeitos**: uma oportunidade para as empresas. Rio de Janeiro: Mauad, 2000. [1] [2]

DÖRRE, J., et al. **Text Mining:** Finding Nuggets in Mountains of Textual Data In Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), pp. 398-401. [1] [2] [3]

FAYYAD, U., et al. **From Data Mining to Knowledge Discovery in Databases.** AAAI/MIT Press, 1997. [1] [2] [3] [4] [5] [6] [7] [8]

FELDMAN, R., et al. **A domain independent environment for creating information extraction modules**, in Proceedings of the tenth international conference on Information and knowledge management, 2001, p586-588, ACM Press. Disponível em: http://doi.acm.org/10.1145/502585.502699. Acesso em: 02 set. 2005. [1] [2]

FRAWLEY W. J.; PIATETSKY-SHAPIRO G.; MATHEUS C. J. **Knowledge discovery in databases:** An overview. In G. Piatetsky-Shapiro and W. J. Frawley, editors, Knowledge Discovery in Databases, pages 1--27. AAAI/MIT Press, 1992. [1] [2] [3] [4] [5]

HAN J.; KAMBER M. **Data Mining**: Concepts and Techniques, Simon Fraser University, Morgan Kaufmann Publishers, 2000. [1] [2]

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**, The MIT Press, 2001, 546 p. [1] [2]

HEARST, M.A. 1999. **Untangling text data mining**. In Proceedings of ACL'99: the 37th annual meeting of the association for computational linguistics, University of Maryland, June 20–26 (invited paper). Disponivel em: <http://www.ai.mit.edu/people/jimmylin/papers/Hearst99a.pdf.>. Acesso em 19 jan. 2006. v [1]

KANTARDZIC, M. **Data Mining**: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, 2003, 343 p. [1]

KAO, A.; POTEET, S. **Text Mining and natural Language Processing** – Introduction for the Special Issue, SIGKDD Explorations, v.7, Issue 1, 2005. [1]

KHATTREE, R. ; NAIK, D. N. **Multivariate Data Reduction and Discrimination with SAS® software,** Cary, NC: SAS institute Inc, John Wiley, 2000. [1] [2]

KODRATOFF Y. **Knowledge Discovery in Texts:** A Definition, and Applications. in Foundation of Intelligent Systems, Ras & Skowron (Eds.) LNAI 1609, Springer 1999. [1] [2] [3]

KONCHADY, M.; **Text Mining Aplication Programming**; Charles River Media, Boston, Massachussets, 2006. [1]

KOTLER, P. A generic concept of marketing, **Journal of Marketing,** v.36, p. 31 – 42, April 1972. [1] [2]

LIMA, G. Â. B.; Interfaces entre a ciência da informação e a ciência cognitiva; **Ciência da Informação**., Brasília, v. 32, n. 1, p. 77-87, jan./abr. 2003. [1] [2]

LYMAN. P; VARIAN, H. R. **How Much Information 2003?** School of Information Management and Systems, University of California at Berkeley. 2003. Disponível em: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/. Acesso em 02 set 2005. [1]

LOH, S.; WIVES, L. K.; OLIVEIRA, José Palazzo Moreira de. **Descoberta Proativa de Conhecimento em Textos**: Aplicações em Inteligência Competitiva. In: International Symposium on Knowledge Management/Document Management (ISKM/DM 2000), III, Nov, 2000. Proceedings.... Curitiba/PR: PUC-PR, 2000. p.125-147. [1]

LUHN, H. P. **The automatic creation of literature abstracts**. IBM Journal, 159–165. 1958. [1] [2] [3] [4]

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. MIT Press, 1999. [1] [2] [3] [4] [5] [6]

MIRANDA, A. **Ciência da Informação**: teoria e metodologia de uma área em expansão; Elmira Simeão (Organizadora); Brasília; Thesaurus, 2003. [1] [2]

MOENS, M. F. **Automatic indexing and abstracting of document texts**. Massachusetts: Kluwer Academic Publishers, 2000. [1] [2]

MORESI, E. A. D.; Delineando o valor do sistema de informação de uma organização. **Ciência da Informação**, Brasília, v. 29, n.1, 2000. [1]

PEPPERS, D. ROGERS, M. **Gerente Um a Um**. Rio de Janeiro: Campus, 2000. [1]

RAJMAN, M.; BESANÇON, R. **Text Mining**: Natural Language techniques and Text Mining applications. Chapman & Hall, 1997. [1]

ROBREDO, J. **Da Ciência da informação revisitada aos sistemas humanos de informação**. Brasília: Thesaurus; SSRR Informações, 2003, 262 p. [1]

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da informação**, vol 24, número 1, 1995. [1] [2]

SHARMA, S. **Applied Multivariate Techniques**. John Wiley & Sons, Inc. 1996. [1] [2]

SHELTMAN, S. Muito além do 0800. **Revista Propaganda** nº 52. p. 24-26, nov 1999. [1] [2]

STONER, J. A. F; FREEMAN, R. Edward. **Administração**, Rio de Janeiro: PHB, 1995, 533 p. [1]

TAN, A.-H. Text mining**:** The state of the art and the challenges. **In Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases**, p. 65–70, 1999. [1] [2] [3] [4]

TERRA, E. **Curso Prático de Gramática**, São Paulo, Scipione, 2002, 423 p [1]

TRYBULA, W. J. Text mining. **Annual Review of Information Science and Technology**, vol. 34, 1999, p. 385-419. [1] [2] [3] [4]

VAN RIJSBERGEN, C. J. **Information Retrieval**, 2nd edition. Dept. of Computer Science, University of Glasgow. 1979. Disponível em: http://www.dcs.gla.ac.uk/Keith/Preface.html. Acesso em 16 set. 2006. [1]

VIEIRA, R.; LIMA, V. L. S. de. **Lingüística computacional:** princípios e aplicações. Disponível em: <http://www.inf.unisinos.br/~renata/laboratorio/publicacoes/jaia12-vf.pdf >. acessado em 02 out. 2005. [1]

WAKEFIELD, T.. **A Perfect Storm is Brewing:** Better Answers are Possible by Incorporating Unstructured Data Analysis Techniques, DM Direct, 2004. Disponível em: < http://www.datawarehouse.com/article/?articleid=4766>. Acesso em 01 set. 2006. [1]

WEISS, S. M., et al. **Text Mining**: Predictive Methods for Analyzing Unstructured Information. Springer, New York, 2005. [1] [2] [3] [4] [5]

WIVES, L. K.. **Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. 2001. Exame de Qualificação (doutorado em Ciência da Computação) -- Instituto de Informática, UFRGS, Porto Alegre. [1]

WIVES, L. K.; LOH, S. **Tecnologias de descoberta de conhecimento em informações textuais** (ênfase em agrupamento de informações). In: OFICINA DE INTELIGÊNCIA ARTIFICIAL (OIA), III, 1999, Tutorial, Pelotas, RS. Proceedings... Pelotas: EDUCAT, 1999. p. 28-48. [1] [2]

WOODFIELD, T. **Mining Textual Data Using SAS® Text Miner for SAS®9 Course Notes**, SAS INSTITUTE INC. Cary, North Carolina, USA. 2004. [1] [2] [3] [4] [5] [6]

YANG Y.; PEDERSEN J.P. A Comparative Study on Feature Selection in Text Categorization. **In Proceedings of the Fourteenth International Conference on Machine Learning** (ICML'97). 1997. [1] [2]

ZIPF, G. K.. **Human Behavior and the Principle of the Least Effort**. Cambridge, MA: Addison-Wesley, 1949. [1]

ZÜLZKE, M. L.. **Abrindo a Empresa para o Consumidor**: A importância de um Canal de Atendimento. 2. ed. Rio de Janeiro: Qualitymark, 1997. [1] [2]