MÁRCIO MARQUES FERREIRA

ESTUDO E APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO EM TEXTO PARA CLASSIFICAÇÃO DE FATURA DE SERVIÇOS DE TELECOMUNICAÇÕES

Monografia apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de *MBA* em Tecnologia de Software

Área de concentração:
Sistemas de Informação

Orientador:
Prof. Dr. Jorge Rady de Almeida Jr.

São Paulo
2008

FICHA CATALOGRÁFICA

Ferreira, Márcio Marques
Estudo e aplicação de técnicas de mineração em texto para classificação de fatura de serviços de telecomunicações / M.M. Ferreira. -- São Paulo, 2008.
74 p.

Monografia (MBA em Tecnologia de Software) – Escola Politécnica da Universidade de São Paulo. Programa de Educação Continuada em Engenharia.

1.Ciência da computação 2.Inteligência artificial I.Universidade de São Paulo. Escola Politécnica. Programa de Educação Continuada em Engenharia II.t.

DEDICATÓRIA

À minha mãe e à Susi.

OK

AGRADECIMENTOS

Agradeço ao professor Jorge Rady, pela orientação e pelo constante estímulo transmitido durante todo o trabalho. Aos amigos, companheiros e mentores do Grupo Telefônica, sem os quais esta obra não seria possível.

À minha mãe que em sua grande sabedoria me guiou pelo caminho da verdade, justiça e honestidade. Ao meu irmão Marciano e à minha tia Esmeralda que me apoiaram em minha educação. Ao meu amigão e compadre Rodrigo. Ao amigo e mentor Araújo, pelos conselhos e apoio nos piores momentos. Ao amigo Fábio Zani, pela orientação em minha carreira. Aos amigos Manoel e Vanderlita, que me apresentaram ao mar e a outras belezas da vida. E à minha pequena Susi, que sempre esteve ao meu lado incentivando e apoiando.

Profecias solenes são, obviamente, um
procedimento fútil, a não ser por fazer
nossos descendentes rirem.

J.B. Priestly

RESUMO

A privatização das telecomunicações no Brasil trouxe inúmeros benefícios para os consumidores. A oferta de novos produtos e serviços, a competição entre as diversas operadoras de telecomunicações e o papel das agências reguladoras deu maior poder aos consumidores, tanto para escolher seu provedor de serviços quanto para exigir os seus direitos. Diante deste cenário, as operadoras de telecomunicações devem zelar em todos os aspectos do negócio para garantir a lucratividade e sustentabilidade. Um dos processos mais críticos é o faturamento dos serviços prestados, no qual a operadora identifica todos os serviços consumidos pelo cliente, aplica as devidas tarifas e emite uma fatura contra o cliente, o qual irá pagá-la ou não mediante ao seu julgamento quanto à exatidão dos produtos/serviços cobrados. Como forma de mitigar este risco, a empresa foco deste trabalho, uma empresa do grupo Telefônica, mantém um processo para localização, análise e correção pontual das contas telefônicas antes da postagem aos clientes. A presente obra tem por objetivo investigar técnicas de mineração em texto para classificação de Notas Fiscais Fatura de Serviços de Telecomunicações em formato digital, implementadas pela tecnologia *AFP® - Advanced Function Printing* (uma tecnologia proprietária da IBM), com o intuito de apoiar o processo da operadora na etapa de localização das contas telefônicas.

Palavras-chave: mineração em texto, *text mining*, classificação de texto, telecomunicações, *data mining*

ABSTRACT

The privatization of telecommunications in Brazil has brought many benefits to consumers. The offering of new products and services, the competition between the various telecom players and the role of regulatory agencies gave more power to consumers, both to choose their service provider as well as to demand their rights. In this scenario, the telecommunications companies must ensure good quality in all aspects of their processes to ensure both profitability and sustainability. The billing is one of the most critical process, in which the company identifies all the services due by the customer, apply the appropriate charges and issue an invoice to the customer, which he/she will pay it or not, upon his own assessment on the accuracy of the products/services billed. As a way to mitigate this risk, a Telefônica group's company (the company focus of this work), maintains a process for searching, analyzing and correcting bills before they are post to customers. This work aims to investigate techniques of text mining for classification of telecom bills in a digital format, implemented by the AFP ® technology - Advanced Function Printing (a proprietary technology from IBM), in order to support the company's process of searching bills.

Keywords: text mining, text classification, telecommunications, data mining

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 Introdução

1.1 Justificativa

Com as privatizações no setor de telecomunicações no Brasil, diversas empresas passaram a operar sob regime de concessão neste setor. A demanda reprimida por produtos e serviços de telecomunicações por parte dos consumidores, e, as exigências contratuais junto ao órgão regulador (Agência Nacional de Telecomunicações – ANATEL), impulsionaram grandes investimentos por parte das empresas operadoras de telecomunicações (infra-estrutura de redes, processos, pessoas, tecnologias etc.), tanto para atender a regulamentações contratuais, como para ganhar competitividade face à crescente oferta de produtos concorrentes e entrada de novos *players* no mercado.

À medida que as empresas aumentavam suas carteiras de clientes, oferecendo uma vasta gama de produtos, serviços, promoções, vantagens etc., cada qual com suas especificidades técnicas (do ponto de vista do produto) e comerciais (do ponto de vista do relacionamento do cliente com a operadora), o processo de faturamento (*billing*) passou cada vez mais a ser um processo crítico dentro da operadora. Com um volume de informações da ordem de milhões de clientes e bilhões de chamadas (com tendência de crescimento), e, não obstante, imposições legais da agência reguladora, o menor erro pode tomar proporções catastróficas. Neste sentido, é imprescindível a presença de fortes pontos de controle no processo de faturamento.

Basicamente o processo de faturamento de serviços de telecomunicações consiste em captar todos os encargos devidos pelo cliente (referente ao consumo e solicitações), tarifá-los e emitir documento de faturamento/cobrança dos serviços contra o cliente. O documento de faturamento/cobrança dos serviços de telecomunicações é a "hora da verdade" para com o cliente, pois o cliente acredita, a priori, que todos os produtos/serviços mencionados no documento para pagamento são devidos por ele efetivamente, e que todos foram tarifados corretamente. Neste

aspecto, a agência nacional de telecomunicações determina que, para cada mil documentos para pagamentos emitidos, apenas dois poderão ter produtos/serviços impugnados por questões de erros de faturamento, sob pena de processo administrativo[1], ou seja, de cada mil clientes que receberam uma conta telefônica apenas dois poderão reclamar sobre problemas na fatura sem que a empresa seja penalizada.

Com este cenário, é imprescindível que uma empresa operadora de telecomunicações possua mecanismos de validação e verificação nas contas telefônicas antes postá-las ao cliente. Uma vez que o volume de informações é demasiado grande (milhares de contas, ainda que se trate de amostragens), o processo de validação e verificação precisa ser ágil para tratar o maior número de contas num intervalo de tempo, e, flexível o suficiente para incorporar modificações.

O processo que inspirou o desenvolvimento da presente obra é a atividade de análise de contas em formato digital, do processo de faturamento de uma empresa operadora de telecomunicações do grupo Telefônica. Regularmente, os analistas de faturamento desta operadora analisam cópias fidedignas das contas telefônicas em formato digital, antes do envio do cliente. Esta atividade tem por objetivo identificar e corrigir problemas pontuais no tocante aos produtos e serviços que serão faturados, bem como quaisquer aspectos na fatura que possam causar uma *percepção* de erro por parte do cliente.

Um dos principais problemas enfrentados pelos analistas de faturamento é a morosidade na localização de contas telefônicas com determinados produtos, serviços, tipos de contas telefônicas e outros aspectos pontuais. Atualmente, estima-se que um analista gaste, em média, 2 horas nesse tipo de localização, considerando todo o seu tempo disponível para análise. Neste sentido, a presente obra procura pesquisar a literatura existente sobre o tema de mineração de textos com o intuito de propor uma solução para o problema enfrentado pelos analistas.

---

[1] Processo administrativo é uma autuação efetuada pela agência nacional de telecomunicações contra a empresa operadora de telecomunicações, em casos que a agência entende que a operadora cometeu alguma infração

1.2 Objetivo

O objetivo deste trabalho é, a partir do estudo de técnicas de mineração em texto, propor uma forma de localização e classificação de contas telefônicas em formato digital mais eficiente, de forma a obter um ganho de desempenho tanto para a companhia como para os analistas.

Conforme mencionado na justificativa (item 1.1), estima-se que, do total de tempo disponível para análise de contas em formato digital, há um gasto médio de 2 horas em pesquisas para identificação de tipos de contas, produtos e serviços. Se houver uma forma de reduzir esse tempo de pesquisa, certamente haverá ganhos comensuráveis (redução de custos) e incomensuráveis (desempenho de atividades mais nobres), tanto para a companhia com para os analistas de faturamento.

1.3 Estrutura do Trabalho

No capítulo 2 apresenta-se o ambiente corporativo que motivou a proposta de trabalho, detalhado-se o processo de faturamento como um todo, bem como a atividade de análise de contas telefônicas em formato digital.

No capítulo 3 são apresentados alguns dos principais mecanismos classificação de textos, visualização e edição de taxonomias, identificação de sentenças comparativas, visualização de conjuntos de textos etc.

No capítulo 4 são apresentados dois exemplos de classificação das contas telefônicas em formato digital, procurando alinhar a prática com os conceitos teóricos aprendidos.

No capítulo 5 são apresentadas conclusões para o trabalho, bem como sugestões para trabalhos futuros.

2 Ambiente Corporativo

O ambiente corporativo que motivou a presente pesquisa é a operação de faturamento (*billing*) de uma operadora de telecomunicações do Grupo Telefônica.

2.1 Sobre o grupo

A Telefônica é uma das empresas líderes mundiais no setor das telecomunicações. Está presente na Europa, América Latina e África, atendendo aproximadamente 195,9 milhões de clientes. (TELEFÔNICA, S.A.)

Presente na América Latina há 15 anos, acumula mais de 70 bilhões de euros com investimentos em infra-estrutura e aquisições. E líder na América Latina e em países ibero-americanos. (IBIDEM)

> Desde junho de 2006, a Telefônica é a operadora líder no Brasil, Argentina, Chile e Peru. Também está desenvolvendo importantes operações na Colômbia, Equador, El Salvador, Guatemala, México, Marrocos, Nicarágua, Panamá, Porto Rico, Uruguai e Venezuela. (IBIDEM)

O Grupo Telefônica ocupa a sexta posição no ranking Eurostoxx 50 (20 de novembro de 2006), com mais de 100 bilhões de dólares de capitalização na bolsa. Em junho de 2006, o grupo empregava, em média, 225.879 funcionários (IBIDEM).

No Brasil, o grupo é o maior conglomerado empresarial privado não-financeiro, com receitas de R$ 28 bilhões (em 2004), encerando o ano de 2005 atendendo aproximadamente 44,4 milhões de clientes e empregando diretamente 63 mil pessoas. O Grupo atua no mercado brasileiro de telefonia fixa e celular (por meio de uma *Join Venture* com a Portugal Telecom), acesso à internet (Terra Lycos) e serviços de *call center* e *contact center* (Atento) (TELEFONICA BRASIL).

2.2 Sobre a operação de faturamento

Nesta sessão são apresentados alguns conceitos do ciclo de faturamento de uma empresa operadora de telecomunicações do grupo Telefônica. Aspectos regulatórios (os quais possuem relação aos serviços de telecomunicações que podem ser prestados pela operadora) e tecnológicos (os quais possuem relação com a forma de medição e tarifação dos serviços) foram desconsiderados para efeitos didáticos. Ambos os aspectos são regulados pela Agência Nacional de Telecomunicações (ANATEL).

Os tópicos a seguir apresentam uma visão geral e bastante simplista (uma vez que o processo com um todo é demasiado complexo e foge ao escopo deste trabalho) dos processos que ocorrem desde o consumo do serviço até a emissão da fatura para pagamento.

A
Figura 1, procura fornecer um percurso cognitivo sob o processo de faturamento.

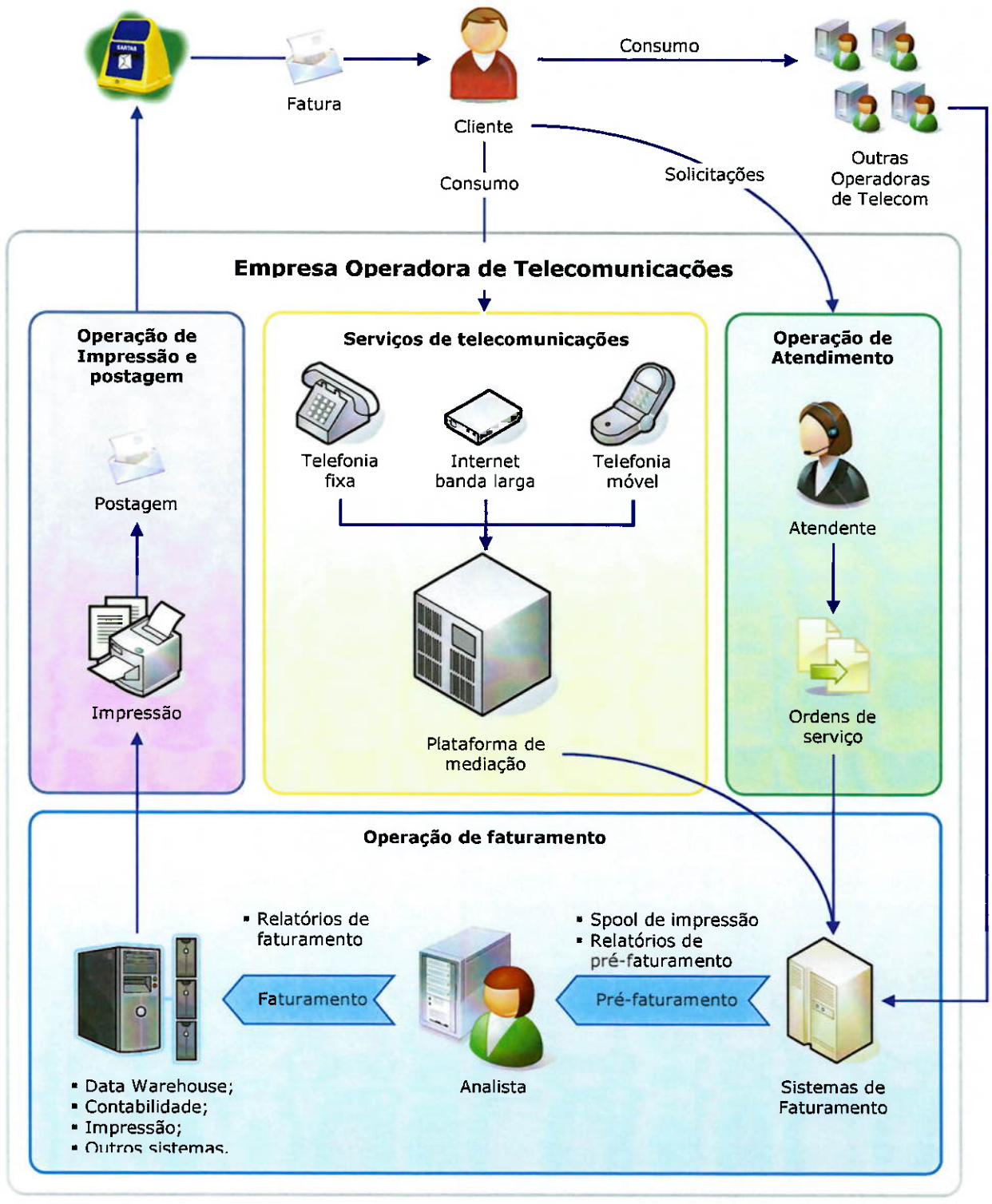


Figura 1 – Visão geral do ciclo do faturamento de serviços de telecomunicações

2.2.1 Faturamento de serviços medidos/bilhetados

A operação de faturamento de serviços medidos/bilhetados é composta pelas seguintes etapas:

1. O cliente consome serviços de telecomunicações da empresa operadora de serviços de telecomunicações (doravante EOT) com a qual contratou os serviços. Esses serviços incluem, todavia não se limitam a: telefonia fixa, telefonia móvel, internet banda larga etc.;

2. Os serviços de telecomunicações consumidos (telefonia fixa, móvel, comunicação de dados etc.) são medidos por uma plataforma mediadora, a qual registra a data da chamada, o horário de início e término, o telefone de origem e destino da chamada (e muitas outras informações);

3. Os sistemas de faturamento recebem as chamadas (bilhetes) da plataforma mediadora, aplicam as regras de tarifação, sumarizam todo o consumo de tráfego do cliente e enviam as informações para consolidação da Nota Fiscal Fatura de Serviços de Telecomunicações (NFFST);

2.2.2 Faturamento de serviços aperiódicos

A operação de faturamento de serviços aperiódicos é composta pelas seguintes etapas:

1. As solicitações de produtos/serviços (ordens de serviço) pelo cliente são registradas pela operação de atendimento da EOT, junto aos sistemas de faturamento;

2. A partir das ordens de serviço, os sistemas de faturamento identificam os produtos/serviços que o cliente adquiriu/cancelou, calcula as tarifas aplicáveis e envia as informações para consolidação da NFFST;

2.2.3 Co-faturamento (*Co-billing*)

A operação de co-faturamento de serviços de telecomunicações de outras operadoras é composta das seguintes etapas:

1. O cliente consume serviços de outra EOT, como por exemplo, realização de chamadas nas modalidades longa distância nacional ou internacional;

2. Os serviços de telecomunicações consumidos (telefonia fixa, móvel, de longa distância nacional/internacional) são medidos por uma plataforma mediadora, a qual registra a data da chamada, o horário de início e término, o terminal de origem e destino da chamada (e muitas outras informações);

3. A EOT responsável pelo registro da chamada aplica as tarifas de acordo com o tipo de chamada e as envia para EOT com a qual o cliente possui contrato, para o devido faturamento;

4. Os sistemas de faturamento da EOT com a qual o cliente possui contrato recebem as chamadas e as envia para consolidação da NFFST;

2.2.4 Atividades de Pré-faturamento

Após término do processamento, os sistemas de faturamento geram diversos arquivos para validação dos analistas de faturamento. Esses arquivos são relatórios de pré-faturamento dos módulos de tráfego, serviços recorrentes, aperiódicos e de terceiros, e, um *spool* de impressão (ver Figura 2) com amostras fidedignas das contas telefônicas (NFFST) que serão impressas e postadas ao cliente.

Basicamente as validações realizadas nos relatórios de pré-faturamento consistem de batimentos e análise de variabilidade dos produtos e serviços de telecomunicações.

A etapa de pré-faturamento é útil para correções pontuais e lançamentos de última hora, os quais são realizados diretamente na conta telefônica do(s) cliente(s).

2.2.5 Atividades de Faturamento

Com exceção da atividade de análise imagem das contas, as atividades da etapa de faturamento são as mesmas da etapa de pré-faturamento, isto é, batimentos e análise de variabilidade dos produtos e serviços.

Ao término da etapa de faturamento são realizadas interfaces com diversos sistemas, como por exemplo, contabilidade, livros fiscais, *Data Warehouse*, impressão das contas etc..

2.3 Sobre a atividade de análise das contas telefônicas em formato digital

A atividade foco deste trabalho é análise do *spool* de impressão das contas telefônicas. O spool de impressão é um arquivo binário, cujo conteúdo consiste de objetos de textos, gráficos e imagens, implementados pela tecnologia *Advanced Function Printing (AFP®)*, um produto da IBM® para impressão de mistos de textos, imagens e gráficos em ambientes de impressão em larga escala. O *spool* de impressão contém amostras fidedignas das contas telefônicas que serão impressas e enviadas aos clientes. A Figura 2 ilustra o *spool* de impressão.

Esta atividade é realizada dez vezes ao mês, em média a cada três dias - uma vez para cada vencimento de fatura (vencimento é um dia fixo no mês que o cliente escolhe para pagamento da fatura) - com um prazo máximo de 8hs para conclusão.

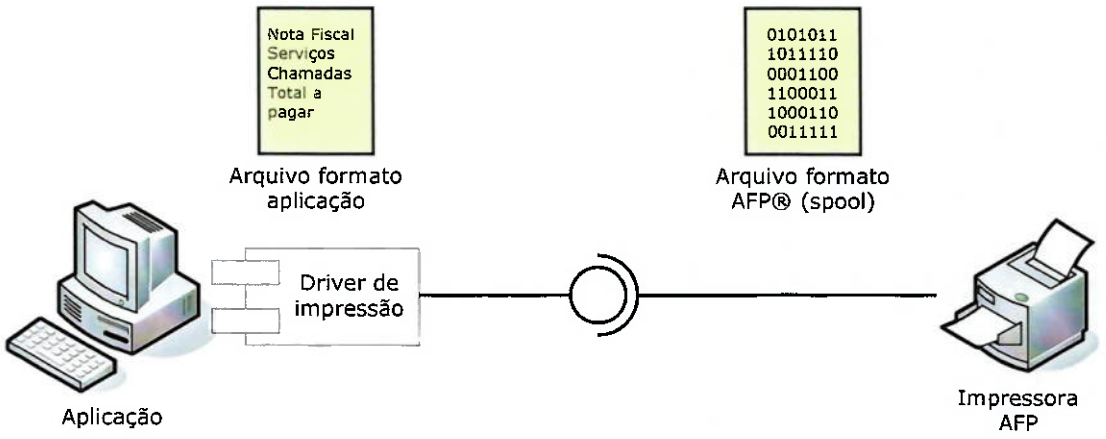


Figura 2 – Spool de impressão no formato AFP®.

A atividade é dividida entre analistas de faturamento especializados e generalistas. Os analistas especializados analisam serviços de tráfego, mensalidades, serviços aperiódicos/eventuais e serviços de terceiros, e, os generalistas analisam a conta como um todo. As análises especializadas consistem em:

- **Serviços de Tráfego**

✓ As chamadas foram tarifadas corretamente de acordo com o plano que o cliente contratou?

✓ Constam todas as informações para entendimento do cliente (data, telefone, localidade, horário de início, modalidade e valor)?

✓ As chamadas estão sendo apresentadas no grupo correto (i.e. chamadas para celular, à cobrar, longa distância etc.)?

- **Serviços Mensais, Aperiódicos e de Terceiros**

✓ Os serviços foram tarifados corretamente? De acordo com a tarifa aplicável? Pro rateados de acordo com a data de solicitação de adesão ou cancelamento do cliente?

- ✓ Constam todas as informações para entendimento do cliente (descrição inteligível do produto/serviço, data de aquisição do produto, período de prestação do serviço, valor do produto/serviço)?

- ✓ Os produtos/serviços estão sendo apresentados no grupo correto (i.e. mensalidades, descontos, serviços de terceiros etc.)?

- **Análise Geral da Conta Telefônica**

- ✓ As totalizações estão corretas?

- ✓ Análise de contas com uma única página;

- ✓ Análise de contas cujo valor total esteja entre R$ 0,01 e R$ 10,00;

- ✓ Análise de contas cujo valor total seja exatamente R$ 0,00;

- ✓ Análise de contas cujo valor total seja menos de R$ 0,00;

- ✓ Análise de contas totalizadoras de contas não pagas: 1, 2, 3 ou 4 contas;

2.4 Definição do Problema

Para efetuar as análises mencionadas no item 2.3, os analistas de faturamento utilizam um navegador padrão Internet Explorer ® com um componente acoplado da IBM (*AFP Viewer*) para visualização de arquivos no formato AFP. Cada arquivo possui em média 500 MB de tamanho, mais de 57.000 páginas, aproximadamente 19.000 termos distintos, e, aproximadamente 14,5 milhões de seqüências de caracteres separadas por espaço (incluindo termos e números e excluindo seqüências de um único caractere).

Seguindo o modelo de lista de verificação apresentado no anexo A, os analistas inserem, na ferramenta supracitada, seqüências de caracteres correspondentes aos

produtos/serviços que irão analisar. Alguns produtos/serviços são facilmente identificados por essa técnica de busca - quando possuem uma seqüência de caracteres específica associada aos mesmos numa única linha -, outros produtos/serviços dependem da sorte para serem identificados – quando possuem duas ou mais seqüências de caracteres independentes da linha em que aparecem. A Figura 3 ilustra essa situação. Suponha que um dos itens da lista de verificação do analista seja analisar uma nota fiscal de outra operadora, a qual tenha faturado uma chamada de longa distância no horário reduzido. Uma vez que a ferramenta de busca disponível limita-se a uma cadeia de caracteres, o usuário deverá identificar primeiramente a operadora desejada (palavra em destaque na figura) e depender da sorte para localizar o restante das informações.

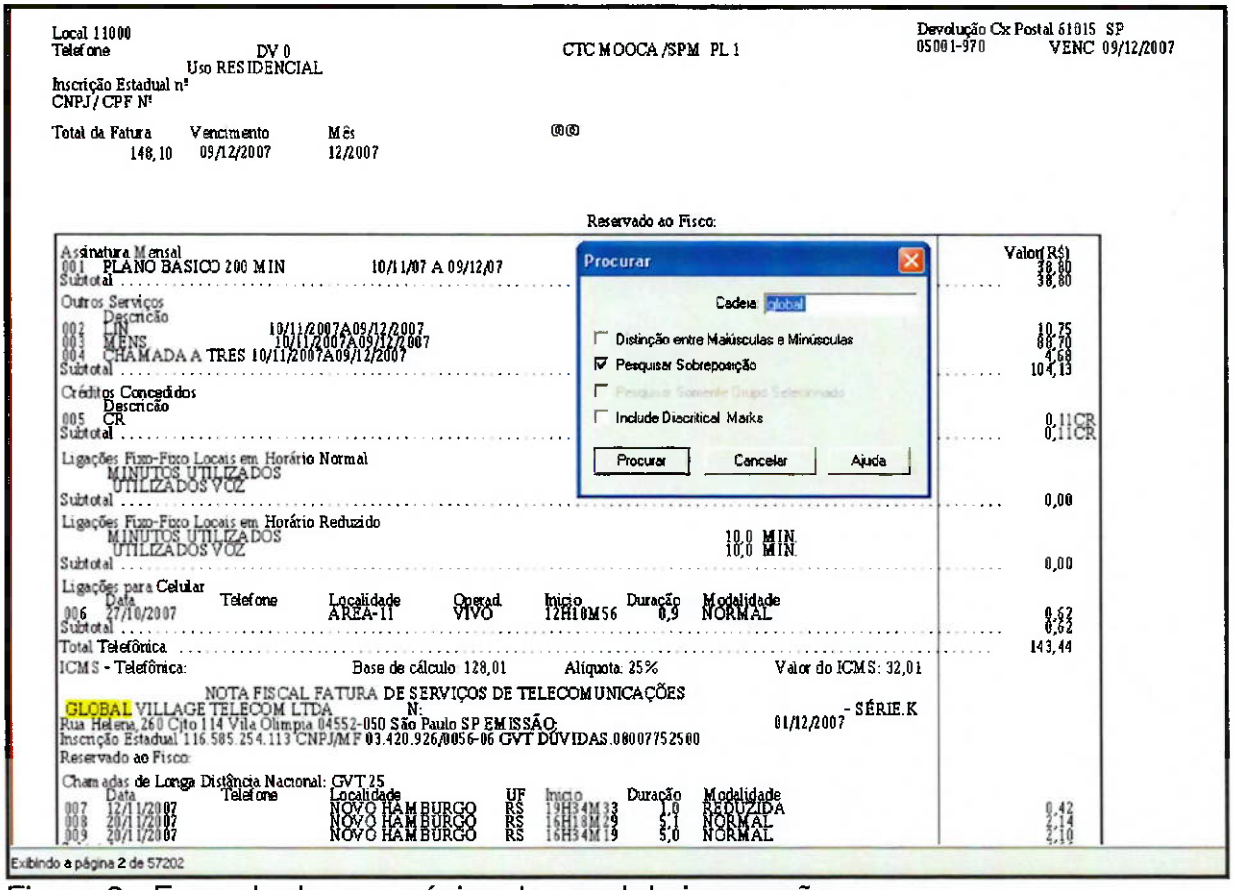


Figura 3 - Exemplo de uma página do spool de impressão

Conforme mencionado no item 2.3, os analistas tem até 8 horas para analisar todos os itens da lista de verificação no spool de impressão. Por experiência dos próprios analistas, estima-se que um único analista analise em média 100 contas num período de 8 horas, com um gasto médio de 2 horas no tipo de busca supracitada,

ou seja, 25% do tempo total é gasto em buscas (considerando-se as buscas bem e mal sucedidas). Descontando-se 2 horas do tempo gasto em buscas, um analista consegue analisar 16 contas por hora (100 contas divididas por 6 horas).

Os números apresentados acima abrem margem para algumas indagações interessantes, tais como: Quanto custa o tempo perdido em pesquisas? Qual o ganho operacional se esse tempo fosse reduzido em 50%? Haveria algum ganho intangível nessa redução?

A primeira questão é facilmente respondida pela aritmética: conhecendo o processo e valor hora/homem da atividade obtém-se o custo da atividade. Já as demais questões são respondidas pelo presente trabalho.

O presente trabalho tem por objetivo apresentar uma solução para os problemas expostos, por meio da pesquisa de métodos e técnicas de mineração em textos.

3 Algoritmos e Técnicas de Mineração em Texto

Segundo Witten e Frank (2005), mineração de dados é um processo que consiste em procurar padrões nos dados. Da mesma forma, "mineração de texto" é um processo que consiste em procurar padrões no texto: mineração de texto "é o processo de se analisar textos com o intuito de extrair informações úteis para algum propósito". Na visão de Witten e Frank (2005), há uma similaridade superficial entre mineração de dados e mineração de textos que escamoteia a real diferença entre ambos. A mineração de dados em um banco de dados relacional, por exemplo, procura informações desconhecidas e potencialmente úteis, implícitas nos dados ali armazenados. Já no caso da mineração de texto as informações estão claramente explicitas e declaradas no corpo do texto.

Do ponto de vista prático, a mineração de texto procura extrair informações úteis do texto para consumo de computadores ou das pessoas, as quais não têm o tempo necessário para ler o texto na íntegra.

Witten e Frank (2005) apontam que a informação minerada deve ser potencialmente útil - um requisito comum tanto para mineração de dados como para mineração de textos na visão desses autores. Potencialmente útil, no sentido de ser capaz de promover algum tipo de ação automática. No que diz respeito à mineração de dados, essa "ação automática" pode ser interpretada de uma forma relativamente independente do domínio de aplicação:

a. Predições não triviais podem ser realizadas na mesma fonte de dados;

b. O desempenho pode ser mensurado através da contagem de sucessos e falhas;

c. Técnicas estatísticas podem ser aplicadas para comparar diferentes métodos de mineração de dados sobre o mesmo problema.

Comparado aos dados utilizados em mineração de dados, textos são, em sua maioria, não-estruturados, amorfos e difíceis de lidar. Não obstante, é difícil mensurar qual tipo "ação automática" pode se obter de mineração de textos de uma forma independente do domínio de aplicação. Nesse sentido, Witten e Frank (2005) apontam um problema que pode ser abordado através da mineração de texto: a classificação de documentos.

3.1 Modelos Bayesianos para Classificação de Documentos

Um domínio de aplicação de mineração de textos é a classificação de documentos, onde cada instância representa um documento e, a classe, o tópico do documento (WITTEN e FRANK, 2005). Uma vez que os documentos são caracterizados por suas palavras, é factível classificá-los através de aprendizagem de máquina considerando-se a presença ou ausência de cada uma de suas palavras como um atributo Booleano. Uma técnica bastante popular para este tipo de aplicação é a Naïve Bayes por ser bastante rápida e precisa (WITTEN e FRANK, 2005).

É importante lembrar que a técnica de Naïve Bayes não considera a freqüência de cada palavra para a determinação da categoria de um documento, ao invés, considera o documento como um "saco-de-palavras" (*Bag of words*). Uma versão modificada da técnica para acomodar a freqüência de cada palavra (conhecida como *multinominal* Naïve Bayes) é apresentada a seguir:

Seja $n_1, n_2, ..., n_k$ o número de vezes que a palavra i ocorra em um documento e, P_1, $P_2, ..., P_k$ sejam as probabilidades de se obter a palavra i de uma amostra do universo de documentos classificados sob a categoria H. Assumindo-se que a probabilidade seja independente do contexto da palavra e de sua posição no documento, essas assertivas levam a *distribuição multinomial* para probabilidades de documentos.

Para esta distribuição, a probabilidade de um documento E condicionada a classe H é:

$$\Pr[E \mid H] \approx N! \times \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_i!}$$

Onde $N = n_1 + n_2 + ... + n_k$ é número de palavras no documento. Explica-se a existência das operações fatoriais pelo fato de que a ordem da ocorrência de cada uma das palavras é imaterial de acordo como o modelo "saco-de-palavras" (*Bag of words*). P_i é estimada computando a freqüência da palavra i no texto de aprendizagem pertencente a H.

Um exemplo pode melhorar a compreensão. Suponha que haja apenas as palavras *amarelo* e *azul* no vocabulário e, um documento classificado na categoria H tenha $\Pr[amarelo|H]=75\%$ e $\Pr[azul|H]=25\%$ (pode-se chamar a classe H de documentos *verdes amarelados*). Suponha que E seja um documento composto pelas palavras *azul amarelo azul* com um tamanho $N = 3$ palavras. Há quatro possibilidades de "sacos-de-palavras", compostos por três palavras. Uma é {*amarelo amarelo amarelo*} e, sua probabilidade de acordo com a fórmula precedente é:

$$\Pr[\{amarelo\, amarelo\, amarelo | H\}] \approx 3! \times \frac{0.75^3}{3!} \times \frac{0.25^0}{0!} = \frac{27}{64}$$

As outras três, com suas respectivas probabilidades são:

$$\Pr[\{azul\, azul\, azul | H\}] = \frac{1}{64}$$

$$\Pr[\{amarelo\, amarelo\, azul | H\}] = \frac{27}{64}$$

$$\Pr[\{amarelo\, azul\, azul | H\}] = \frac{9}{64}$$

Neste exemplo, E corresponde ao último caso (importante lembrar que no modelo "saco-de-palavras" a ordem dos termos é irrelevante), e, sua probabilidade de ser classificado como um documento *verde amarelado* é 9/64, ou 14%.

3.2 Representação de textos para mineração

A dificuldade encontrada para se construir um sistema de classificação de documentos é fortemente influenciada pela linguagem de indexação utilizada para representar o texto (LEWIS, 1992). Apté et al. (1994) concordam com Lewis ao afirmarem,

> ... uma vez que os métodos de aprendizagem de máquina resolvem problemas por meio da análise de amostras descritas através de suas métricas e atributos, os documentos devem ser transformados a priori para classificação através de métodos de aprendizagem de máquina. (APTÉ et al., 1994, p. 3)

Segundo Lewis (1992), a forma mais simples de abordar esse problema é tratar cada palavra do texto como um atributo. Todavia, na visão desse autor, as palavras carregam algumas propriedades intrínsecas, tais como sinônimos, homônimos e polinômios, propriedades que podem gerar problemas durante a classificação. Apté et al. (1994) complementam que,

> Sistemas de busca de documentos são supostos a selecionar documentos sobre algum conceito de interesse do receptor. Entretanto, documentos não possuem conceitos, ao invés, documentos possuem palavras, e. palavras não correspondem diretamente a conceitos. (APTÉ et al., 1994, p. 3)

Para os seres humanos, é de relativa facilidade estabelecer relações entre conceitos e palavras. A título de exemplo, é trivial para um ser humano adulto a correlação entre o conceito *médico* e a palavra *cardiologista*, tal como *banco* enquanto instituição financeira e *banco* enquanto objeto para sentar-se. Segundo Apté et al. (1994) isso se dá pelo fato dos seres humanos acumularem ao longo de sua vida um vasto conhecimento da gramática de uma língua e do mundo de uma forma geral, os seres humanos adquirem, numa única palavra, cultura.

Em um extenso trabalho de investigação sobre aprendizagem de máquina na categorização de textos, Sebastiani (2002) corrobora Lewis e Apté et al. quando afirma que textos não podem ser diretamente interpretados por classificadores. Contudo, no que diz respeito à forma de representação de textos para classificação, Sebastiani condiciona a escolha da representação a uma definição clara do que

deve ser considerado como uma unidade de texto significativa, bem como quais são regras semânticas significativas de combinação dessas unidades. Sebastiani pondera que este último aspecto muitas vezes é desconsiderado e um texto d_j é representado num espaço vetorial de termos ponderados $\vec{d}_j = \langle w_{1j}, ..., w_{|T|j} \rangle$, onde T é um conjunto de termos (muitas vezes também denominados características ou atributos), os quais ocorrem ao menos em um documento, e $0 \leq w_{kj} \leq 1$ representa quanto o termo t_k contribui para a semântica do documento d_j. (SALTON e BUCKLEY, 1988)

Nas sessões a seguir são apresentados alguns modelos de classificação de documentos pesquisados pelo autor desta obra.

3.3 Categorização de Textos Através de Aprendizagem de Máquina

Na década de 80, a abordagem mais popular para criação de classificadores consistia em construir manualmente, através de técnicas de engenharia do conhecimento (*knowledge engineering*), um sistema capaz de tomar decisões sobre categorização de textos. Tal sistema consistia num conjunto de regras lógicas definidas manualmente para cada categoria, do tipo (SEBASTIANI, 2002):

if (*fórmula de Forma Disjuntiva Normal*) **then** (*categoria*)

Exemplo:

If	((*trigo* ∧ *fazenda*	**or**
	(*trigo* ∧ *commodity*)	**or**
	(*bushels* ∧ *exportação*)	**or**
	(*trigo* ∧ *toneladas*)	**or**
	(*trigo* ∧ *inverno* ∧ ¬ *macio*))	**then** TRIGO **else** ¬TRIGO

Figura 4 - Exemplo de regra

Essa técnica de classificação define que, uma categoria *<categoria>* será atribuída a um documento se ao menos uma das proposições da *fórmula de Forma Disjuntiva Normal* for satisfeita.

A desvantagem na utilização dessa abordagem reside na (a) falta de portabilidade do método e (b) no gargalo de aquisição de conhecimento:

a. A utilização do classificador em um domínio diferente daquele para o qual fora criado implicará na a reconstrução de todas as regras;

b. A criação de regras deve ser elaborada por um "engenheiro do conhecimento" (*knowledge engineer*) auxiliado por um especialista no domínio do problema. Não obstante, a adição de novas regras requer a intervenção conjunta desses profissionais; (SEBASTIANI, 2002).

No início da década de 90 os métodos de aprendizagem de máquina começam a ganhar popularidade. Nessa abordagem um classificador para a categoria c_i é automaticamente construído através de um processo de indução, o qual observa as características de um conjunto de documentos previamente classificados sob a categoria c_i ou $'c_i$ (leia-se complemento de c_i) por um especialista no domínio do problema. A partir das características observadas, o processo de indução "escolhe minuciosamente" as características de um documento não classificado para determinar se o mesmo deve ser classificado sob a categoria c_i ou $'c_i$. (SEBASTIANI, 2002).

Segundo Sebastiani (2002), a vantagem desse método sobre a engenharia de conhecimento reside no esforço para a criação de um construtor automático de classificadores ao invés de um classificador. Se tal construtor estiver disponível tudo que se necessita é de um conjunto de documentos previamente classificados, eliminando os problemas de adição de novas regras e portabilidade do classificador.

Desta forma, depreende-se que o recurso mais importante na abordagem de aprendizagem de máquina são os documentos previamente classificados. Numa situação real, o caso mais favorável à aplicação dessa abordagem seria o da

organização que, em algum momento do passado, empregou a classificação manual de documentos e decidiu automatizar o processo. O caso menos favorável seria o da organização que não dispõe desses documentos previamente classificados (SEBASTIANI, 2002). Sebastiani defende que, ainda que tais documentos não estejam disponíveis, a abordagem de aprendizagem de máquina é preferível à engenharia de conhecimento:

> Na verdade, é mais fácil classificar manualmente um conjunto de documentos do que construir e ajustar um conjunto de regras, uma vez que é mais fácil caracterizar um conceito explicitamente (i.e. selecionar instâncias de documentos) do que implicitamente (i.e. descrever conceitos em palavras ou descrever um procedimento de reconhecimento de suas instâncias). (SEBASTIANI, 2002, p. 9)

Segundo Sebastiani (2002) a abordagem de aprendizagem de máquina está condicionada à disponibilidade de um corpo inicial $\Omega = \{d_1,\ldots,d_{|\Omega|}\} \subset D$ de documentos pré-classificados sob a categoria $C = \{c_1,\ldots,c_{|C|}\}$. Isto é, o valor da função $\breve{\Phi} : D \times C \rightarrow \{T, F\}$ (aproxima-se de *VERDADEIRO–TRUE* – ou *FALSO-FALSE*) é conhecido para cada par $\langle d_j, c_i \rangle \in \Omega \times C$ (SEBASTIANI, 2002). Segundo o autor, um documento d_j é um exemplo positivo da classe c_i se $\breve{\Phi}(d_j, c_i) = T$ e um exemplo negativo se $\breve{\Phi}(d_j, c_i) = F$.

Sebastiani (2002) afirma que tanto em aplicações práticas como acadêmicas, é comum avaliar a efetividade de um classificador Φ após sua construção. Nesse sentido, o corpo inicial Ω é dividido em duas partes (não necessariamente iguais):

- A primeira parte consiste de um conjunto de dados para treinamento e validação do classificador: $TV = \{d_1,\ldots,d_{|TV|}\}$. O classificador Φ para as categorias $C = \{c_1,\ldots,c_{|C|}\}$ é construído por indução, através da observação de características dos documentos (SABASTIANI, 2002);

- A segunda parte consiste de um conjunto de dados para testar a efetividade do classificador: $Te = \{d_{|TV|+1},\ldots,d_{\Omega}\}$. Cada documento $d_j \in Te$ é alimentado no classificador Φ, e suas decisões Φ (d_j, c_i) são comparadas às decisões do

especialista no domínio do problema $\breve{\Phi}(d_j, c_i)$. A medida de efetividade é baseada na quantidade de classificações coincidentes efetuadas pelo classificador em relação às classificações efetuadas pelo especialista. (IBIDEM)

Sebastiani (2002) alerta, no entanto, para o fato de que os documentos Te de testes não podem participar de forma alguma do processo de indução para criação dos classificadores. Segundo Mitchell (1996, apud SEBASTIANI, 2002), tal prática resultaria em resultados experimentais irreais e sem caráter científico.

3.4 Métodos Interativos de Classificação de Documentos

Spangler e Kreulen (2002) apresentam uma abordagem orientada à edição e validação de taxonomias para classificação de documentos. Segundo os autores, uma vez que o conhecimento disponível em textos supera a quantidade de informações disponíveis em dados brutos, torna-se um desafio tirar proveito dessa riqueza de informações (SPANGLER e KREULEN, 2002). A solução encontrada por esses autores para superar esse desafio consiste em visualizar a coleção de documentos agrupados em taxonomias: "Taxonomia é a organização hierárquica natural da informação, alinhada com as metas de negócios, a organização e os processos." (SPANGLER e KREULEN, 2002)

A Figura 5 mostra um exemplo de taxonomia. Considere, a título de exemplo, um conjunto de documentos com uma breve descrição das 500 maiores companhias da Revista Fortune. Cada documento pode descrever o que a companhia produz, onde está localizada sua sede, em quais tecnologias são líderes, quais parcerias foram firmadas etc.. Cada uma destas taxonomias provê uma forma de definir o significado de "similaridade". No contexto hipotético exposto acima, similar no tocante à geografia significa estarem presentes na mesma região geográfica, enquanto em tecnologia significa serem competidoras no mesmo mercado (SPANGLER et al, 2002).

Geografia
América do Norte
Europa
África
Ásia
...

Tecnologia
Engenharia Química
Redes de computadores
Combustíveis alternativos
Software
...

Indústria
Energia
Transporte
Computadores
Serviços
...

Figura 5 - Exemplo de Taxonomia (SPANGLER et al, 2002)

A abordagem de Spangler e Kreulen (2002) baseia-se em três passos: visualização, edição e validação de resultados de *clusters*. Para tanto, esses autores utilizam uma ferramenta denominada *eClassifier*, a qual viabiliza a aplicação dessa abordagem.

3.4.1 Visualizando a Taxonomia

Na visão de Spangler e Kreulen (2002), antes que os usuários possam editar uma taxonomia eles precisam conhecer as categorias em que os documentos estão classificados, bem como seus relacionamentos:

- Os textos são representados em um espaço vetorial, através das freqüências ponderadas dos atributos (palavras e frases);
- Cada categoria é representada por um centróide[2];
- Utilização da métrica de similaridade de *cosine*[3] para comparar as distâncias entre os documentos e os centróides.

[2] Centróide é um ponto cujas coordenadas são as médias das coordenadas dos pontos de uma figura geométrica; centro geométrico.
[3] Uma medida de similaridade entre vetores. Muito popular em *clustering* de textos.

3.4.1.1 Sumários

Spangler e Kreulen (2002) apresentam duas técnicas interessantes para sumarizar os documentos pertencentes a uma categoria, pois, "não podemos esperar que os usuários gastem seu tempo lendo os documentos de cada categoria". A primeira técnica consiste em obter um gráfico de barras, com duas barras: (1) a primeira barra representa a percentagem de documentos na categoria que possuem o atributo; (2) a segunda barra representa freqüência do atributo no restante dos documentos; as barras são ordenadas em ordem decrescente da diferença entre (1) e (2), mostrando ao usuário os atributos mais importantes de uma categoria.

A segunda técnica consiste na criação de uma árvore de decisão que descreve quais combinações de atributos definem a categoria.

3.4.1.2 Visualização

A estratégia utilizada para mostrar ao usuário uma ou mais categorias relacionadas às taxonomias consiste em exibir o vetor de atributos dos documentos como pontos no espaço. O resultado será a apresentação de agrupamentos de documentos com palavras similares. O problema desta estratégia é que o espaço dimensional dos documentos é muito grande (milhares de termos) para ser apresentado num plano cartesiano (SPANGLER & KREULEN, 2002). A abordagem utilizada pelos autores para reduzir a dimensionalidade dos documentos, procurando reter o máximo de informações relevantes, foi a aplicação do método CViz (DHILLON et al., 1998) o qual sustenta-se nas centróides de três categorias para definir o plano mais interessante para projeção dos documentos enquanto pontos neste plano.

3.4.2 Editando a Taxonomia

Spangler e Kreulen (2002) defendem que o próximo passo após visualizar e compreender o conteúdo de uma taxonomia consiste na edição e modificação da mesma. A recomendação dos autores é de não procurar uma taxonomia perfeita, a qual cubra todas as lacunas possíveis (na visão dos mesmos é improvável que tal taxonomia exista), mas sim permitir que o especialista no domínio do problema decida qual a melhor taxonomia sob seu ponto de vista. Nesse sentido, a ferramenta utilizada pelos autores permite a edição da taxonomia no tocante às categorias e aos documentos.

- Edição da taxonomia no tocante às categorias: o usuário pode excluir uma categoria, unir duas ou mais categorias e mover uma categoria nos diferentes ramos da árvore de categorias.

- Edição da taxonomia no tocante aos documentos: o usuário pode selecionar um ou mais documentos e mudá-los de categoria (ferramentas de busca por palavra-chave são disponibilizadas ao usuário para facilitar a seleção de documentos), excluí-los de uma categoria ou até copiá-los para outra categoria (neste último caso um único documento pode coexistir em duas categorias);

3.4.3 Validando a Taxonomia

O último aspecto da abordagem de Spangler e Kreulen (2002) consiste na validação da taxonomia. Os autores propõem duas formas de validação da taxonomia. A primeira consiste na inspeção visual das categorias, procurando validar se cada categoria contém os documentos relacionados àquela categoria. Uma outra forma de validação visual é através dos gráficos de clusters dos documentos: as áreas de sobreposição às margens dos clusters são os primeiros candidatos à inspeção. A segunda forma de inspeção consiste na aplicação de métricas utilizadas para validação de resultados de algoritmos para criação de clusters. Spangler e Kreulen

(2002) alertam que, apesar de tais métricas não serem totalmente aplicáveis a taxonomias que foram modificadas para incorporar conhecimentos de domínio específico, alguns de seus conceitos podem ser aproveitados. Os autores aplicaram duas métricas de validação para as taxonomias: coesão e distinção.

- Coesão: uma medida de similaridade da categoria. Trata-se da distância média de *cosine* dos documentos na categoria em relação ao centróide da categoria.

- Distinção: uma medida de distinção entre categorias. Trata-se da diferença entre um menos a distância de *cosine* de uma categoria em relação ao centróide da categoria vizinha mais próxima. (SPANGLER & KREULEN; 2002)

3.5 Uma Ferramenta para Visualização e Compreensão de uma Coleção de Documentos

Spangler et al. (2002) apresentam um sistema chamado *MindMap* para exploração e navegação de conceitos e tópicos de uma coleção de documentos, permitindo o uso de múltiplas taxonomias, visualização e interação com o usuário. Segundo os autores, a novidade que o *MindMap* traz é a uma interface modelada em técnicas utilizadas em *brainstorming*.

Tal como apresentado no item 3.4, os documentos precisam ser classificados em múltiplas taxonomias antes que o usuário possa navegar e explorar a coleção de documentos. Segundo Spangler et al. (2002), muitas abordagens podem ser utilizadas para gerar múltiplas taxonomias sobre uma coleção de documentos. A abordagem utilizada pelos autores foi:

- Aplicação de algoritmos de *clustering* sobre o domínio de termos de uma coleção de documentos;

- Eliminação de palavras freqüentes (*stopwords*) e palavras que não agregam valor ao conteúdo;

- Representação dos documentos através de um vetor de freqüências ponderadas dos termos restantes;

3.5.1 Visualização

Após a criação de múltiplas taxonomias o próximo passo (e o grande desafio) é permitir ao usuário encontrar a categoria (ou categorias) mais relevante ao tópico de seu interesse dentre as diversas taxonomias. Supondo que o usuário tenha interesse no tópico *digital* e realize uma consulta com essa palavra, a ferramenta *MindMap* cria um gráfico radial contendo a consulta ao centro e as classes na periferia, as quais são selecionadas dentre todas as taxonomias que mais se aproximam à consulta. A Figura 6 apresenta um exemplo de uma consulta realizada pelo usuário.

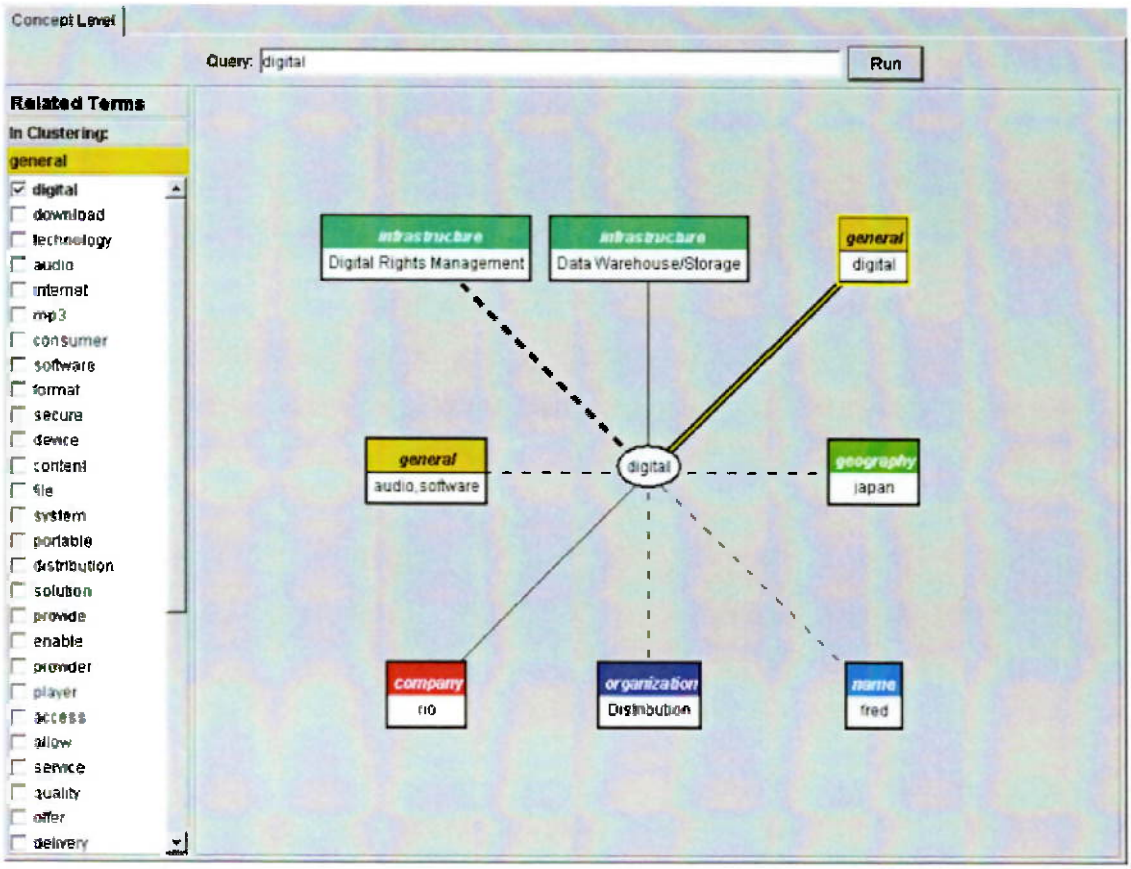


Figura 6 - Gráfico radial (SPANGLER et al, 2002)

Os passos para elaboração do gráfico radial são (SPANGLER et al, 2002):

- A consulta é convertida num modelo de espaço vetorial para seleção das classes;
- O vetor da consulta é comparado com todas as classes de cada taxonomia;
- Primeiramente as taxonomias cujo dicionário possui o maior número de palavras-chave correlatas à consulta são selecionadas;
- Dentre essas taxonomias, são apresentadas as 8 classes cujos centróides estão mais próximos à consulta no modelo de espaço vetorial;
- A variação na espessura e no tipo de linha representa o quanto a taxonomia está relacionada à consulta;

A ferramenta *MindMap* permite ao usuário refinar sua pesquisa selecionando um termo na listagem de termos relacionados (Figura 6) ou por meio da exploração de termos numa árvore binária. Uma vez que o usuário selecione uma classe do gráfico radial, um diagrama de árvore binária é apresentado para que o usuário refine ainda mais sua pesquisa:

- O nó raiz representa a coleção completa de documentos, os quais satisfazem a condição da consulta realizada pelo usuário, na classe selecionada;
- Cada ramo da árvore divide os documentos baseando-se na presença ou ausência de algum termo;
- Inicialmente a árvore é expandida em até três camadas, onde cada ramo é baseado no termo que melhor divide os documentos.
- A
- Figura 7 mostra um exemplo da árvore binária: o nó raiz, taxonomia *company*, é representado pelo termo *rio*, ramifica-se pelos nós *+multimedia* e *–multimedia* (i.e

documentos que contêm o termo *multimedia* – indicado pelo sinal de adição '+' – e documentos que não contém o termo *multimedia* – indicado pelo sinal de subtração '-';

- Cada nó apresenta o número e a percentagem de documentos representados pelo mesmo, juntamente com o termo cuja presença ou ausência o caracteriza. Se o usuário desejar expandir um nó, uma relação dos 5 termos que melhor podem representar o próximo nó é apresentada ao usuário;

Segundo os autores, a vantagem dessa abordagem para refinamento de consultas reside no fato de permitir ao usuário trabalhar com subconjuntos de documentos gerenciáveis do ponto de vista prático. (SPANGLER et al, 2002).

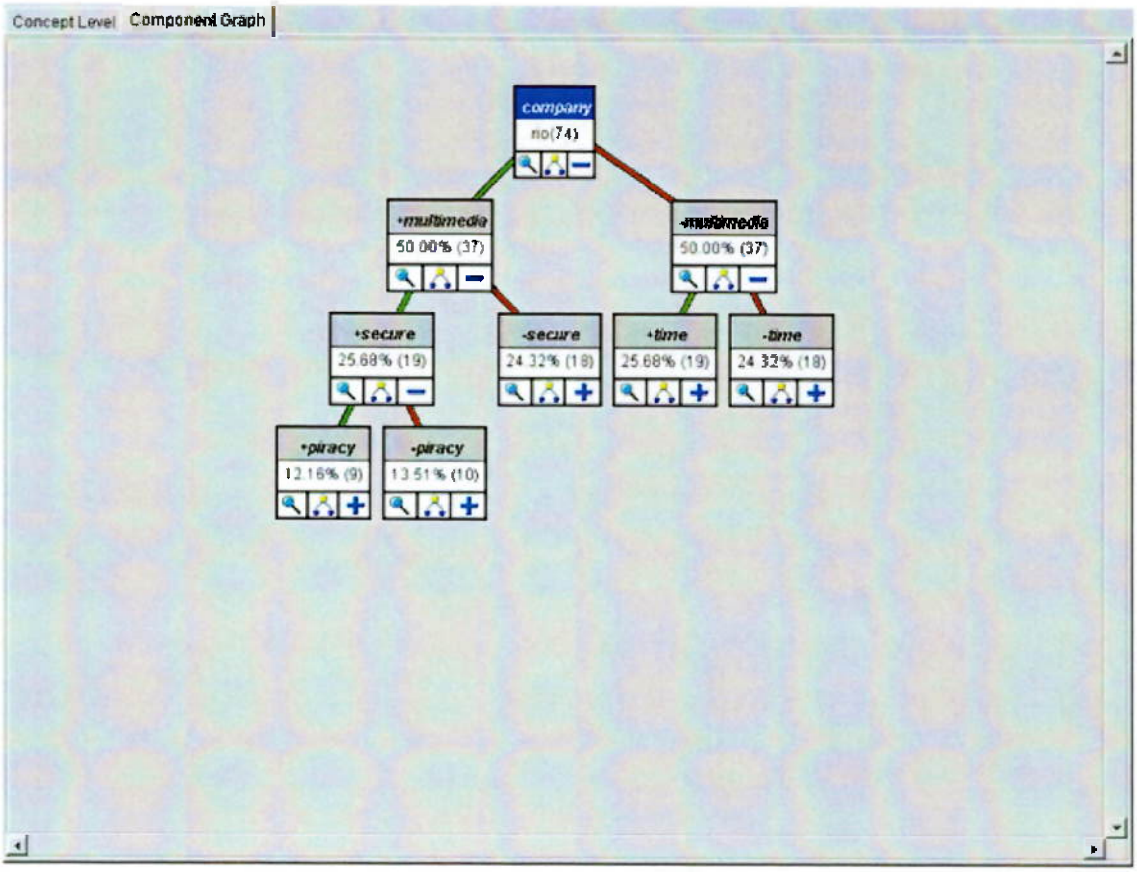


Figura 7 - Arvore Binária (SPANGLER et al, 2002)

O último passo após o usuário selecionar um nó na árvore binária é apresentar os exemplos de documentos que satisfazem à consulta. Segundo Spangler et al (2002), uma forma bastante simples de apresentação é listar os exemplos em ordem

crescente de distância da classe em relação ao centróide. Ponderam, todavia, que essa abordagem reduz o entendimento do usuário a um número.

Uma forma mais elucidativa de apresentação, do ponto de vista prático, seria representar cada documento como um ponto num espaço geométrico dimensional, reutilizando o vetor de freqüências ponderadas dos termos criados anteriormente. Nessa abordagem, a posição do ponto (documento) no espaço é determinada considerando os termos que o documento contém. Spangler et al (2002) alertam para os mesmos problemas relacionados à alta dimensionalidade (uma dimensão para cada termo) dos documentos, abordados no tópico anterior.

> Para contornar esse problema, nós encontramos um plano de projeção "interessante" no espaço de alta dimensionalidade. Projetamos neste plano encontrando a intersecção entre o plano e uma linha normal traçada em direção a cada ponto. Observamos que o plano mais interessante, sob a perspectiva de distinção entre categorias e exemplos, é o plano desenhado pelo centróide de três categorias. (SPANGLER et al, 2002)

Spangler et al (2002) utilizam um sistema de coordenadas baseado em pontos de interesse (Oslen, et al, 1993) para projeção dos pontos.

A ferramenta *MindMap* apresenta no plano três categorias, sendo uma delas a categoria selecionada pelo usuário e, as outras duas, suas vizinhas mais próximas determinadas pela métrica de *cosine* entre os centróides. Os centróides das três categorias definem o plano de visualização. Todos os pontos (documentos) nas três categorias que satisfazem à consulta são apresentados na visualização.

A Figura 8 mostra um exemplo da visualização dos documentos no espaço geométrico.

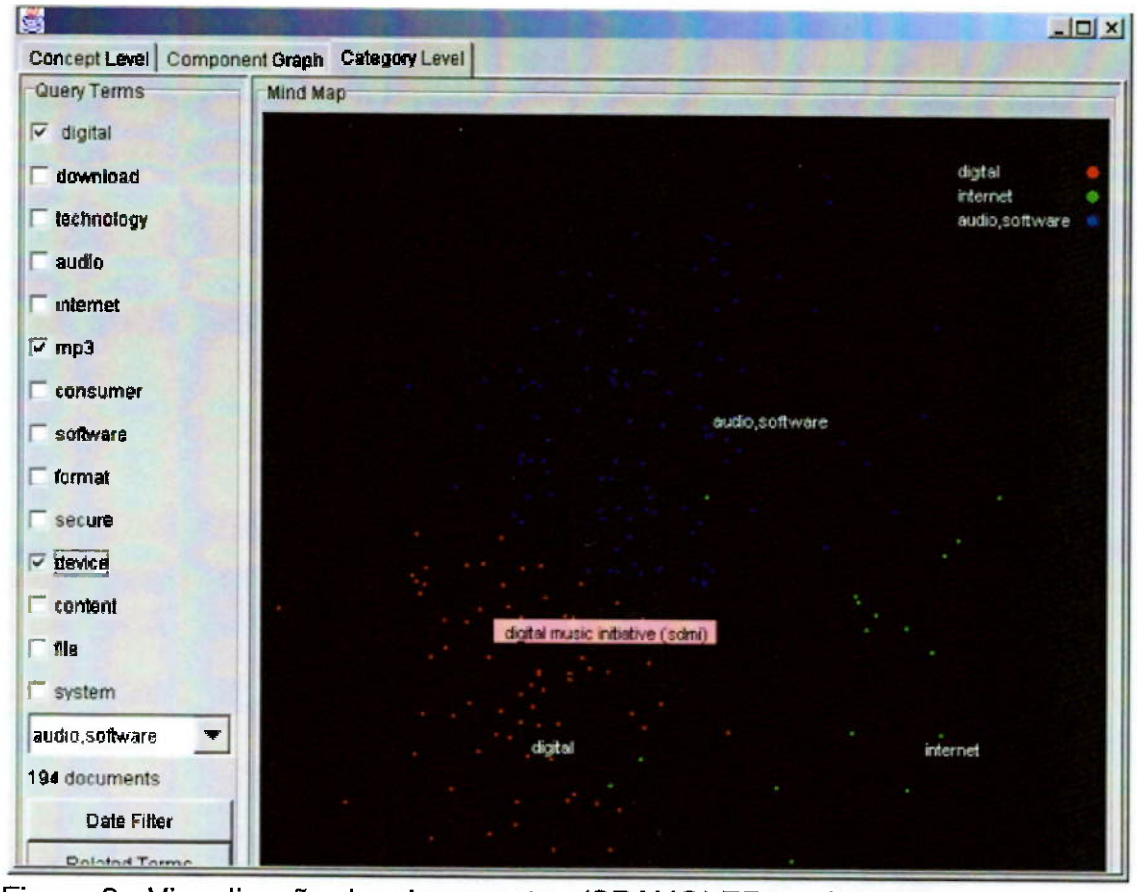


Figura 8 - Visualização dos documentos (SPANGLER et al, 2002)

Segundo Spangler et al (2002) a vantagem dessa abordagem reside no fato de que documentos (pontos) próximos uns dos outros no espaço geométrico também compartilham termos comuns. Isto facilita a identificação de documentos relacionados a um documento de interesso do usuário. Outra vantagem apresentada pelos autores é que, a apresentação de classes relacionadas pode revelar documentos fora da área de interesse do usuário que possam ser relevantes. Segundo os autores, uma fraqueza dessa abordagem reside na visualização de uma taxonomia por vez (SPANGLER et al, 2002).

3.6 Identificação de Sentenças Comparativas em Documentos [4]

Jindal e Bing (2006) apresentam uma técnica para identificar sentenças comparativas em documentos. Os autores salientam que pesquisas anteriores foram conduzidas no sentido de se extrair sentimentos e opiniões de textos, os quais são temas relacionados à sua técnica, porém se diferenciam no sentido de carregarem uma subjetividade intrínseca. Os autores ponderam, todavia, que as sentenças comparativas também podem ser subjetivas ou objetivas. A título de exemplo, tomemos as seguintes sentenças:

"*Car X is very ugly*" – sentença subjetiva (expressa opinião/sentimento);

"*Car X is much better than Car Y*" – sentença comparativa subjetiva;

"*Car X is 2 feet longer than Car Y*" – sentença comparativa objetiva.

Na visão dos autores é bastante perceptível a diferença entre as sentenças que expressam opinião das sentenças comparativas (embora as sentenças comparativas também possam expressar opinião) (JINDAL; BING, 2006).

Antes de discorrer sobre a técnica proposta por Jindal e Bing (2006), é importante delinear o escopo do problema estudado pelos autores. Esses autores apresentam inicialmente uma perspectiva lingüística sobre comparações na língua inglesa, a qual identifica dois grandes grupos de tipos comparativos, os metalingüísticos e os proposicionais (DORAN et al, 1994 apud JINDAL; BING, 2006). Os autores apontam duas limitações na perspectiva lingüística para sentenças comparativas:

1. Sentenças não comparativas com palavras tipicamente comparativas:

---

4 Neste tópico, algumas sentenças foram mantidas em seu idioma original propositalmente, pois, sua tradução para o português implicaria em um estudo mais aprofundado de lingüística da língua portuguesa no sentido de se obter um exemplo equivalente ao dos autores no tocante a sintaxe e semântica de sentenças comparativas, estudo esse que foge ao escopo desta obra.

"*In the context of speed, faster means better*"

"*John has to try his best to win this game*"

Na visão dos autores, embora ambas as sentenças possuam palavras comparativas elas não o são para fins práticos.

2. Sentenças comparativas que não possuem palavras características de comparação:

   "*In market capital, Intel is way ahead of Amd*"

   "*Nokia, Samsung, both cell phones perform badly on heat dissipation index*"

   "*The M7500 earned a World bench score of 85, whereas Asus A3V posted a mark of 89*"

Para contornar a primeira limitação os autores utilizaram métodos de aprendizagem de máquina. Já para a segunda limitação, adicionaram *preferências do usuário* (comparação indireta do tipo "*I prefer Intel than Amd*") e *comparações implícitas* ("*câmera X has 2 MP, whereas câmera Y hás 5 MP*").

Conhecida a perspectiva lingüística e suas limitações (bem como, como contorná-las) Jindal e Bing (2006) definem o escopo do problema como "o estudo das comparações no tocante às sentenças":

Sentença comparativa:

"Uma sentença comparativa é uma sentença que expressa uma relação baseada em similaridades ou diferenças entre mais de um objeto" (JINDAL; BING, 2006).

Objetos e seus atributos:

"Um objeto é uma entidade que pode ser uma pessoa, um produto, uma ação, etc. sob comparação em uma sentença comparativa. Cada objeto tem um conjunto de atributos, os quais são usados para comparar objetos." (ibidem).

Não obstante, Jindal e Bing (2006) agrupam as sentenças comparativas em quatro tipos:

1. **Não-igual graduável:** expressa relações do tipo *menor que* ou *maior que* (relações que implicam numa ordem entre os atributos dos objetos);

2. **Eqüitativa:** expressa relação do tipo *igual a* (relação onde os atributos de dois ou mais objetos são equivalentes);

3. **Superlativo:** expressa relação do tipo *o maior* ou *o menor* (relação onde os atributos de um objeto ocupam uma ordem superior ou inferior sobre todos os outros objetos);

4. **Não-graduável:** sentenças que comparam dois ou mais objetos, mas não estabelecem uma ordem entre os mesmos.

3.6.1 Técnica para Identificação de Sentenças Comparativas

Neste tópico é apresentada uma técnica para identificação de sentenças comparativas.

3.6.1.1 Regras de Classes Seqüenciais

A técnica apresentada por Jindal e Bing (2006) é uma combinação de *mineração de regra de classe seqüencial (class sequencial rule mining* – doravante *CSR)* e

métodos de aprendizagem de máquina. Uma CSR é uma regra com um padrão à esquerda e um rótulo de classe à direita.

Seja $I = \{i_1, i_2, \ldots, i_n\}$ um conjunto de itens. Uma *seqüência* é uma lista ordenada de vários de itens. Um *conjunto de itens* X é um conjunto de itens não-vazio. Uma seqüência s é denotada por $\langle a_1 a_2 \ldots a_r \rangle$, onde a_i é um *conjunto de itens*, também denominado *elemento* de s, e, a_i é denotado por $\{x_1, x_2, \ldots, x_k\}$, onde x_j é um item. Um item pode ocorrer uma única vez num elemento de uma seqüência, porém poderá ocorrer múltiplas vezes em diferentes elementos. Uma seqüência $s_1 = \langle a_1 a_2 \ldots a_r \rangle$ é uma *subseqüência* de outra seqüência $s_2 = \langle b_1 b_2 \ldots b_m \rangle$, se existir inteiros $1 \leq j_1 < j_2 < \ldots < j_{r-1} \leq j_r$ de forma que $a_1 \subseteq b_{j_1}, a_2 \subseteq b_{j_2}, \ldots, a_r \subseteq b_{j_r}$. Também pode-se dizer que s_2 contem s_1 (JINDAL, BING; 2006).

Exemplo:

Seja $I = \{1, 2, 3, 4, 5, 6, 7\}$. A seqüência $\langle \{3\}\{4, 5\} \rangle$ está contida em $\langle \{6\}\{3, 7\}\{4, 5, 6\} \rangle$, uma vez que $\{3\} \subseteq \{3, 7\}$ e $\{4, 5\} \subseteq \{4, 5, 8\}$. Porém, $\langle \{3, 8\} \rangle$ não está contido em $\langle \{3\}\{8\} \rangle$ e vice-versa.

A seqüência de entrada D para mineração é um conjunto de pares: $D = \{(s_1, y_1), (s_2, y_2), \ldots, (s_n, y_n)\}$, onde s_i é uma seqüência e $y_i \in Y$ é sua etiqueta de classe. Y é o conjunto de todas as classes. No contexto da pesquisa de Jindal e Bing, $Y = \{comparativo, não\text{-}comparativo\}$. Uma CSR é uma implicação da forma:

$$X \rightarrow y, \text{ onde } X \text{ é uma seqüência e } y \in Y.$$

Uma instância (s_i, y_i) em D é dita por *cobrir* a CSR se X é uma subseqüência de s_i. Uma instância (s_i, y_i) é dita por *satisfazer* a CSR se X é uma subseqüência de s_i e $y_i = y$. O *suporte* da regra é a fração do total de instâncias em D que a satisfaz, e, a *confidência* da regra é a proporção de instâncias em D que a cobre a e também a satisfaz.

A Tabela 1 fornece um exemplo contendo cinco seqüências de dados e suas respectivas etiquetas de classe c_1 ou c_2. Utilizando um suporte mínimo de 20% e confidência mínima 40%, uma das CSRs descobertas é:

$\langle\{1\}\{3\}\{7,8\}\rangle \rightarrow c_1$ [suporte = 2/5 e confidência = 2/3]

As instâncias 1 e 2 satisfazem à regra, enquanto que as instâncias 1, 2 e 5 cobrem a regra.

#	Seqüência de Dados	Classe
1	$\langle\{1\}\{3\}\{5\}\{7,8,9\}\rangle$	c_1
2	$\langle\{1\}\{3\}\{6\}\{7,8\}\rangle$	c_1
3	$\langle\{1,6\}\{9\}\rangle$	c_2
4	$\langle\{3\}\{5,6\}\rangle$	c_2
5	$\langle\{1\}\{3\}\{4\}\{7,8\}\rangle$	c_2

Tabela 1 - Um exemplo de seqüência de dados para mineração de CSRs (JINDAL, BING; 2006)

3.6.1.2 Construção do Conjunto de Dados para Mineração

Jindal e Bing (2006) defendem que a melhor forma de representar o conjunto de dados para mineração é tratar cada sentença como uma seqüência de palavras. Esses autores ponderam, entretanto, que as palavras não devem ser utilizadas na sua forma in natura para representação da sentença, pois as sentenças podem diferir no conteúdo, porém ser semelhantes no tocante aos padrões lingüísticos subjacentes. Tome-se, por exemplo, as seguintes sentenças, as quais comparam objetos completamente diferentes (ibidem):

"*Intel is better than Amd*", e

"*Laptops are smaller than desktop PCs*"

Jindal e Bing afirmam que a utilização das palavras em sua forma in natura neste exemplo, não captura qualquer padrão lingüístico, exceto pelo fato de haver a

palavra "*than*" em comum. Todavia, um ser humano pode perceber claramente um padrão.

A abordagem encontrada por Jindal e Bing para tornar esses padrões perceptíveis aos métodos de aprendizagem de máquina, foi a substituição de cada palavra por uma marca. Os autores utilizaram uma ferramenta denominada *Brill's Tagger* (BRILL, 1992), a qual marca cada palavra da sentença com seu respectivo grupo gramatical (*Part of speech*), seguindo as recomendações do Penn Tree Bank Part of Speech Tagging Scheme (SANTORINI, 1990). Os principais grupos gramaticais usados pelos autores foram (marca: grupo gramatical):

- NN: Nome;
- NNP: Nome Próprio;
- VBZ: Verbo, tempo presente, 3ª. pessoa do singular;
- JJ: Adjetivo;
- RB: Advérbio;
- JJR: Adjetivo, comparativo;
- JJS: Adjetivo, superlativo;
- RBR: Advérbio, comparativo;
- RBS: Advérbio, superlativo;
- DT – Determinante;
- CD – Número Cardinal.
- IN – Preposição ou Conjunção subordinativa

Conforme apontam suas pesquisas, Jindal e Bing (2006) afirmam que um pequeno conjunto de palavras-chave identifica a maioria das sentenças comparativas, porém a precisão é baixa. Para contornar o problema de baixa precisão os autores eliminaram todas as sentenças que não possuem palavras-chave, bem como geraram regras de classe seqüencial para filtrar as sentenças não comparativas.

Para seus experimentos, Jindal e Bing (2006) compilaram manualmente uma lista de 30 palavras (palavras terminando em *er, exceed, beat, outperform, etc.*) encontradas em sentenças comparativas e utilizaram um dicionário eletrônico para identificar sinônimos para as mesmas. Após um ajuste manual, chegaram a uma lista de 69 termos. Os autores lembram que algumas sentenças comparativas não-graduáveis não utilizam qualquer das palavras-chaves compiladas. Desta forma adicionaram manualmente palavras como *but, whereas, on the other hand* etc., as quais são comumente utilizadas em comparações. Adicionaram também, as marcas de grupo gramatical que identificam comparações: JJR, RBR, JJS e RBS. Ao final, os autores obtiveram uma lista de 83 palavras-chave e frases-chave (do tipo "Number one" e "up against") (JINDAL; BING, 2006). Formalmente, os autores definem a conjunto de palavras-chave, K, como:

$K =$ { *JJR, RBR, JJS, RBS* } $\cup$
{ palavras como *favor, prefer, win, beat, but,* etc } $\cup$
{ frases como *number one, up against,* etc. }

Jindal e Bing (2006) alertam para o fato de que nem todas as sentenças que contenham essas palavras são comparativas. Em seus experimentos identificaram que apenas 32% das sentenças que possuíam ao menos uma palavra-chave, eram efetivamente sentenças comparativas. Ponderam, entretanto, que as palavras-chaves foram responsáveis por identificar 94% das sentenças comparativas: i.e. 94% de *recall* e 32% de precisão.

Jindal e Bing (2006) tomam os seguintes passos para construção do conjunto de dados para mineração:

1. Para cada sentença que contenha ao menos uma palavra-chave ou frase-chave, compor as seqüências pelos termos que estão a um raio de 3 palavras à frente e atrás de cada palavra-chave ou frase-chave;

2. Cada termo é substituído sua marca do grupo gramatical. As palavras-chave são combinadas com a respectiva marca. O motivo para isso é que algumas palavras-chave podem ter mais de uma marca do grupo gramatical.

3. Uma etiqueta de classe é associada a cada seqüência caso a sentença seja comparativa ou não-comparativa.

Exemplo:

O exemplo a seguir mostra uma sentença comparativa e sua respectiva seqüência que irá compor o conjunto de dados para mineração:

Sentença: *"this/DT camera/NN has/VBZ significantly/RB more/JJR noise/NN at/IN iso/NN 100/CD than/IN the/DT Nikon/NN 4500/CD"*

Seqüência: ⟨{NN}{VBZ}{RB}{moreJJR}{NN}{IN}{NN}⟩ *comparativa*

3.6.1.3 Aprendizagem de Máquina para Classificação

Os experimentos iniciais conduzidos por Jindal e Bing (2006) procuraram identificar e selecionar para cada sentença, as regras que apresentassem o melhor resultado no tocante à confidência. Caso a etiqueta de classe da regra seja *comparativa* a sentença será classificada como tal, caso contrário será classificada como não-comparativa. Os autores apontam, no entanto, que os resultados para essa abordagem não foram satisfatórios (apresenta precisão de apenas 46%). Jindal e Bing acreditam que o principal motivo para esse resultado é que uma determinada sentença pode satisfazer a muitas regras, muitas vezes com classes conflitantes, e, escolher uma das classes arbitrariamente pode ser perigoso (JINDAL; BING, 2006)

A solução encontrada por Jindal e Bing foi utilizar o modelo Bayesiano de classificação, uma vez que é capaz de combinar múltiplas probabilidades para se chegar a uma decisão de classificação baseada em probabilidade (JINDAL; BING, 2006). Segundo esses autores, o modelo Bayesiano provê uma solução natural para o tipo de problema supracitado, não obstante, os resultados obtidos após a aplicação desse modelo foram muito melhores (precisão de 79%).

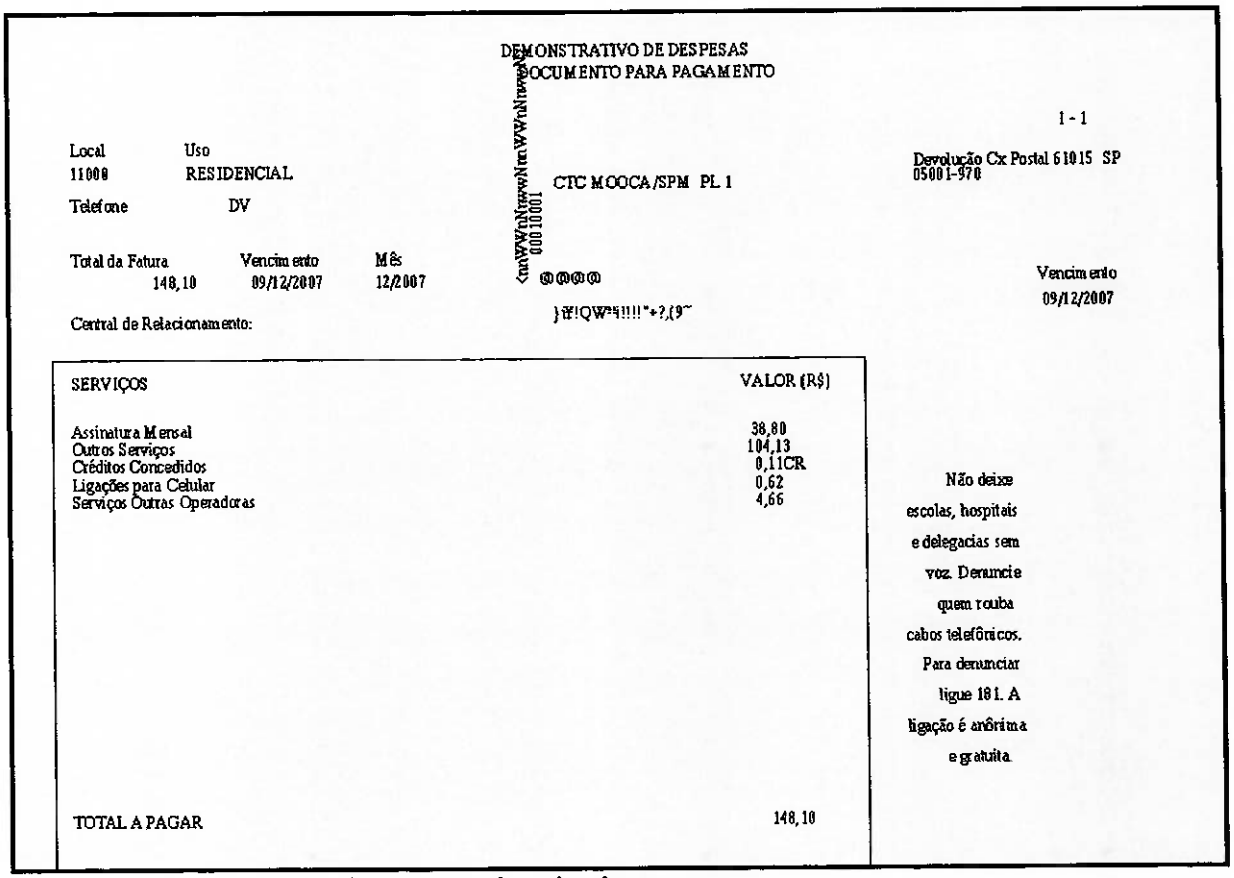
4 Aplicação

Conforme apresentado no capítulo 2, o presente trabalho tem por objetivo apresentar uma solução para o problema de pesquisa textual no *spool* de impressão. No item 2.2, foi abordado que os analistas de faturamento têm um prazo de 8 horas para concluir a análise do *spool* de impressão e gastam, em média, 2 horas em pesquisas textuais para cobrir todos os itens da lista de verificação. Na prática, a "mecânica" do processo é a seguinte: (1) o analista insere uma cadeia de caracteres correspondente a um item da lista de verificação na ferramenta de busca e aguarda a localização do item; (2) se a busca foi bem sucedida, o analista analisa a conta; (3) concluída a análise, o analista e repete a busca para analisar ao menos outra conta semelhante; (4) se a busca foi mal sucedida, o analista tenta o próximo item da lista de verificação; os passos (1), (2), (3) e (4) são repetidos até o fim da lista de verificação. Importante ressaltar que, a estimativa do gasto de 2 horas em pesquisas textuais baseia-se simplesmente na repetição da etapa (1) para cada item da lista de verificação, ou seja, inserir uma cadeia de caracteres para busca e aguardar a resposta da ferramenta. Por meio deste processo, um analista experiente consegue analisar em média 100 contas no prazo estipulado.

Também foi visto no capítulo 3 que, na visão de Witten e Frank (2005), os textos são, em sua maioria, não-estruturados e amorfos para o consumo de computadores. Não obstante, na visão de Sebastiani (2002) os textos precisam ser transformados antes de serem interpretados por classificadores.

A despeito do que afirmam Witten e Frank, o conteúdo do *spool* de impressão pode ser considerado relativamente estruturado. O *spool* de impressão é composto basicamente de três tipos de documentos, os quais são padronizados e possuem delimitações específicas para apresentação de cada tipo de informação:

1. Demonstrativo de Despesas: documento de valor comercial, sem valor fiscal, o qual sumariza todos os gastos com serviços de telecomunicações pelo cliente, e que também serve como boleto de pagamento junto às instituições financeiras. A Figura 9 ilustra um exemplo de demonstrativo: (1) O cabeçalho da página identifica o tipo de documento e, (2) Os serviços são agrupados e sumarizados ao centro da página;



DEMONSTRATIVO DE DESPESAS
DOCUMENTO PARA PAGAMENTO

1 - 1

Local 11000 Uso RESIDENCIAL
Telefone DV

CTC MOOCA/SPM PL 1

Devolução Cx Postal 61015 SP 05001-970

Total da Fatura 148,10 Vencimento 09/12/2007 Mês 12/2007

Vencimento 09/12/2007

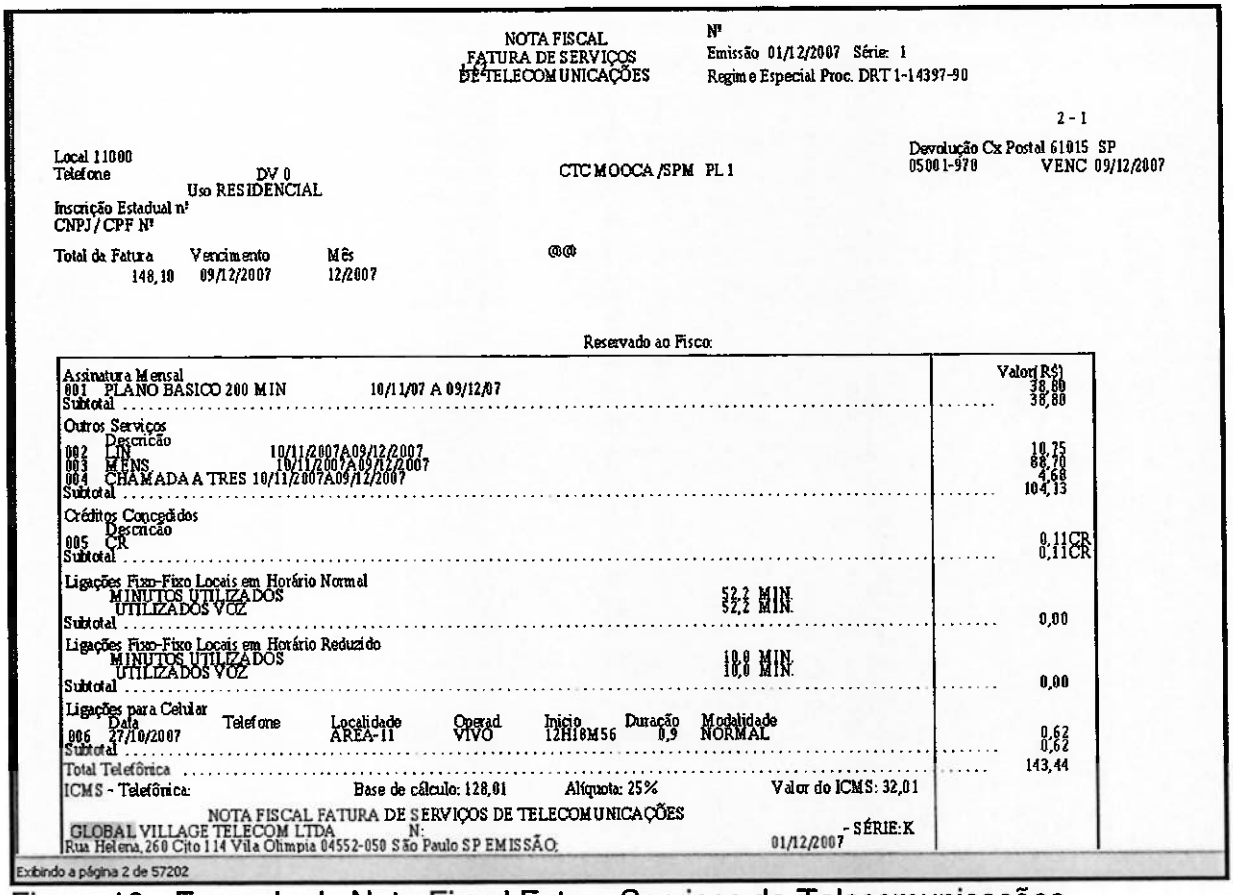
Central de Relacionamento:

SERVIÇOS	VALOR (R$)
Assinatura Mensal	38,80
Outros Serviços	104,13
Créditos Concedidos	0,11CR
Ligações para Celular	0,62
Serviços Outras Operadoras	4,66
TOTAL A PAGAR	148,10

Não deixe escolas, hospitais e delegacias sem voz. Denuncie quem rouba cabos telefônicos. Para denunciar ligue 181. A ligação é anônima e gratuita.

Figura 9 - Exemplo de demonstrativo de despesas

2. Nota Fiscal Fatura de Serviços de Telecomunicações (NFFST): documento de valor fiscal, o qual apresenta detalhadamente todos os serviços de telecomunicações prestados pela empresa operadora de telecomunicações (EOT) juntamente com os impostos. A Figura 10 ilustra um exemplo de nota fiscal. Tal como no demonstrativo de despesas, o cabeçalho da página identifica o tipo de documento, os produtos/serviços consumidos pelo cliente são numerados seqüencialmente e valorados (um produto/serviço por linha);



NOTA FISCAL
FATURA DE SERVIÇOS
DE TELECOMUNICAÇÕES

Nº
Emissão 01/12/2007 Série: 1
Regime Especial Proc. DRT 1-14397-90

2 - 1

Local 11000
Telefone DV 0
Uso RESIDENCIAL
Inscrição Estadual nº
CNPJ / CPF Nº

CTC MOOCA /SPM PL 1

Devolução Cx Postal 61015 SP
05001-970 VENC 09/12/2007

Total da Fatura 148,10 | Vencimento 09/12/2007 | Mês 12/2007

@@

Reservado ao Fisco:

		Valor(R$)
Assinatura Mensal		
001 PLANO BASICO 200 MIN 10/11/07 A 09/12/07		38,80
Subtotal		38,80
Outros Serviços		
Descrição		
002 LIN 10/11/2007A09/12/2007		10,75
003 MENS 10/11/2007A09/12/2007		88,70
004 CHAMADA A TRES 10/11/2007A09/12/2007		4,68
Subtotal		104,13
Créditos Concedidos		
Descrição		
005 CR		0,11CR
Subtotal		0,11CR
Ligações Fixo-Fixo Locais em Horário Normal		
MINUTOS UTILIZADOS	52,2 MIN	
UTILIZADOS VOZ	52,2 MIN	
Subtotal		0,00
Ligações Fixo-Fixo Locais em Horário Reduzido		
MINUTOS UTILIZADOS	10,0 MIN	
UTILIZADOS VOZ	10,0 MIN	
Subtotal		0,00
Ligações para Celular		
Data Telefone Localidade Operad Inicio Duração Modalidade		
006 27/10/2007 AREA-11 VIVO 12H18M56 0,9 NORMAL		0,62
Subtotal		0,62
Total Telefônica		143,44

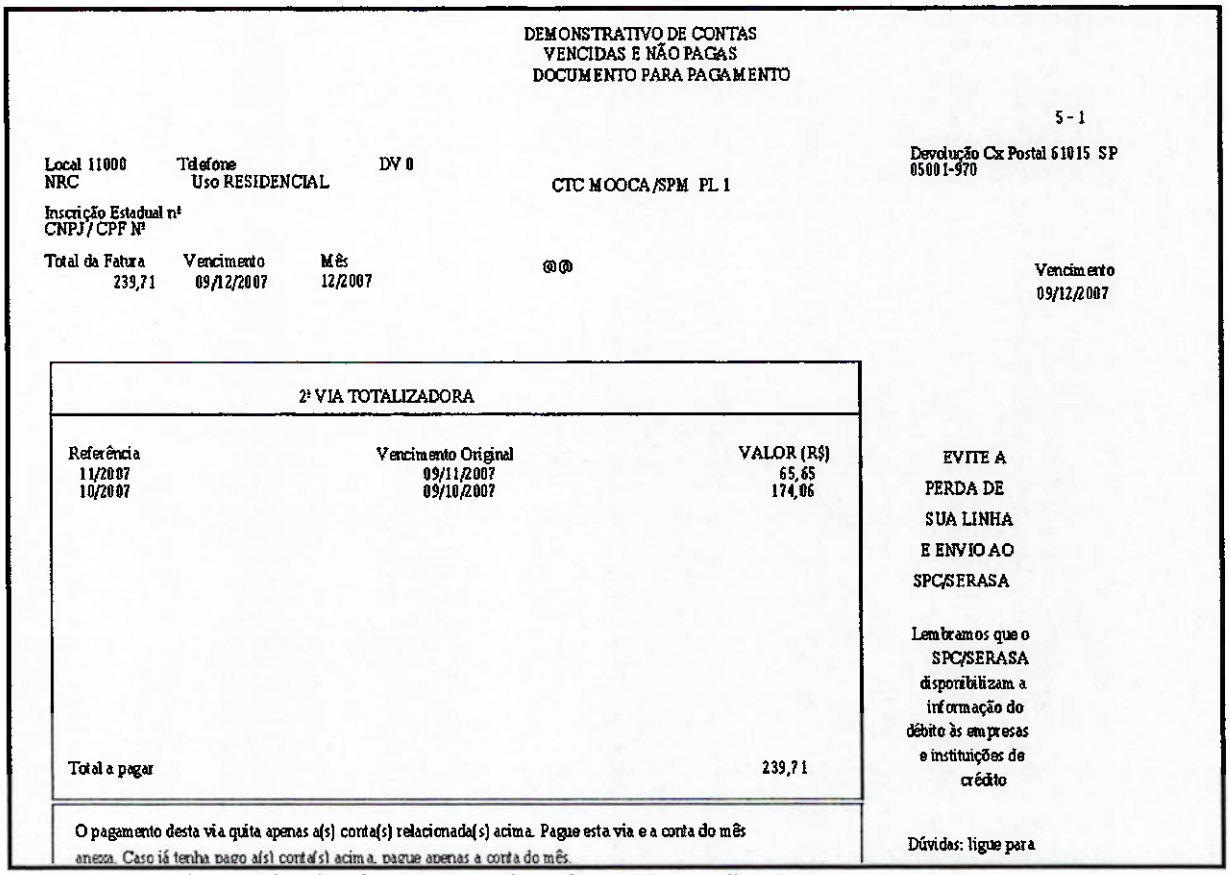
ICMS - Telefônica: Base de cálculo: 128,01 Alíquota: 25% Valor do ICMS: 32,01

NOTA FISCAL FATURA DE SERVIÇOS DE TELECOMUNICAÇÕES
GLOBAL VILLAGE TELECOM LTDA N: - SÉRIE:K
Rua Helena,260 Cjto 114 Vila Olimpia 04552-050 São Paulo SP EMISSÃO: 01/12/2007

Exibindo a página 2 de 57202

Figura 10 - Exemplo de Nota Fiscal Fatura Serviços de Telecomunicações

3. Demonstrativo de contas não pagas: documento de valor comercial, sem valor fiscal, o qual sumariza todas as contas vencidas e não pagas pelo cliente, e que também serve como boleto de pagamento junto às instituições financeiras. A Figura 11 ilustra um exemplo de demonstrativo de contas não agas. Tal como nos documentos apresentados anteriormente, o cabeçalho da página identifica o tipo de documento e, o detalhe da página apresenta o mês de referência das faturas vencidas, o vencimento original e o respectivo valor.



DEMONSTRATIVO DE CONTAS
VENCIDAS E NÃO PAGAS
DOCUMENTO PARA PAGAMENTO

5 - 1

Local 11000 Telefone DV 0
NRC Uso RESIDENCIAL
CTC MOOCA/SPM PL 1

Devolução Cx Postal 61015 SP
05001-970

Inscrição Estadual nº
CNPJ / CPF Nº

Total da Fatura	Vencimento	Mês
239,71	09/12/2007	12/2007

@@

Vencimento
09/12/2007

2ª VIA TOTALIZADORA

Referência	Vencimento Original	VALOR (R$)
11/2007	09/11/2007	65,65
10/2007	09/10/2007	174,06

Total a pagar 239,71

O pagamento desta via quita apenas a(s) conta(s) relacionada(s) acima. Pague esta via e a conta do mês anexa. Caso já tenha pago a(s) conta(s) acima, pague apenas a conta do mês.

EVITE A PERDA DE SUA LINHA E ENVIO AO SPC/SERASA

Lembramos que o SPC/SERASA disponibilizam a informação do débito às empresas e instituições de crédito

Dúvidas: ligue para

Figura 11 - Exemplo de demonstrativo de contas não pagas

É importante lembrar que, tanto o demonstrativo de despesas como a nota fiscal fatura de serviços de telecomunicações possuem objetos de imagem em seu conteúdo, como por exemplo, logotipos e códigos de barra. Neste sentido, é válida a observação de Sebastiani (2002) quanto à transformação dos textos para interpretação dos classificadores. Para os fins práticos deste trabalho, são utilizados apenas os objetos de texto, os quais são classificados por classificadores desenvolvidos *ad hoc*, com o intuito de se obter uma abordagem para o projeto de desenvolvimento de uma ferramenta mais sofisticada.

4.1 Primeiro Exemplo

Apesar de não ter sido abordada na parte teórica desta obra, o primeiro exemplo utilizou os recursos de expressões regulares oferecidos pela linguagem Java™ (versão 5.0) para classificar cada página do *spool* de impressão (vide Figura 2), como uma forma *ad hoc* de otimização das pesquisas dos analistas. Segundo Cornell e Horstman (2005), expressões regulares são utilizadas para especificar padrões de texto, bem como para localizar cadeias de texto que correspondem a determinado padrão.

Por exemplo:

O padrão "`.*pagar.*\d{3,},[0-9][0-9]cr.*`", pode ser compreendido como:

.	o caractere "." (ponto) representa qualquer caractere;
*	o caractere "*" (asterisco) denota quantificação, de forma que *qualquer caractere* pode aparecer zero ou mais vezes;
p	o caractere "p";
a	o caractere "a";
g	o caractere "g";
a	o caractere "a";
r	o caractere "r";
.	o caractere "." (ponto) representa qualquer caractere;
*	o caractere asterisco denota quantificação, de forma que *qualquer caractere* pode aparecer zero ou mais vezes;
\d	um dígito (entre 0 e 9);
{n,}	indica um quantificador, de forma que o caractere ou expressão predecessora ocorra ao menos *n* vezes;
,	o caractere ",";
[0-9]	um dígito entre 0 e 9;
[0-9]	um dígito entre 0 e 9;
c	o caractere "c";
r	o caractere "r";
.	o caractere "." (ponto) representa qualquer caractere;
*	o caractere "*" (asterisco) denota quantificação, de forma que *qualquer caractere* pode aparecer zero ou mais vezes;

Desta forma, a cadeia de caracteres "*pagar 23,42cr*" é uma correspondência bem sucedida do padrão supracitado. Já as cadeias "*pagar 4.123,42cr*" e "*pagar 23,42*" não são. A Tabela 2 apresenta a sintaxe de expressões regulares disponíveis na linguagem Java™ para especificação de padrões de texto.

Para a condução do exemplo, cada página do spool de impressão foi transformada numa página de texto simples, considerando a ordem de apresentação dos termos. Com base na lista de verificação dos analistas de faturamento e, alinhado à sintaxe de expressões regulares da linguagem Java™, foi criado um dicionário de categorias e expressões regulares para utilização no exemplo. A Tabela 3 apresenta algumas das 105 categorias e expressões regulares parametrizadas para o exemplo.

Formalmente, o arquivo transformado D é um arquivo de texto simples composto por páginas de texto, de forma que $D = \{d_1, \ldots, d_{|D|}\}$; onde cada página d é composta por linhas de texto $L = \{l_1, \ldots, l_{|L|}\} \subset d$ e, cada linha é composta de termos de forma que $l_i = \{w_1, \ldots, w_{|l|}\} \subset d$. Seja C um conjunto de pares "categoria/expressão regular" de forma que $C = \{\langle c_1, r_1 \rangle, \ldots, \langle c_{|C|}, r_{|C|} \rangle\}$. Cada página foi classificada por k categorias de forma que $0 \leq k \leq |C|$. A Figura 12 apresenta o algoritmo utilizado para classificação das páginas.

```
for (each p in D)                                          ①
{
        for (each c in C)                                  ②
        {
                if (p matches c.regular expression) then   ③
                        assign c.category to p;
                next c;
        }
        next d;
}
```

Figura 12 – Algoritmo aplicado no primeiro exemplo para classificação das páginas

O algoritmo da Figura 12 deve ser interpretado da seguinte forma:

a) O primeiro laço (①) itera sobre cada página da coleção de documentos;

b) O segundo laço (②) itera sobre cada categoria da coleção de pares categoria-expressão regular, para cada página da coleção;

O teste condicional (③) avalia se a página corresponde à expressão regular daquela iteração. Em caso afirmativo a categoria é atribuída à página, caso contrário a próxima categoria é avaliada.

Sintaxe	**Explicação**
Caracteres	
c	O caractere *c*
\u*nnnn*, \x*nn*, \0*n*, \0*nn*, \0*nnn*	A unidade de código dado o valor hexadecimal ou octal
\t, \n, \r, \f, \a, \e	Os caracteres de controle tabulação, nova linha, retorno de carro, novo formulário, alerta e escape
\c*c*	O caractere de controle correspondente ao caractere *c*
Classes de Caractere	
[C_1C_2...]	Qualquer um dos caracteres representados por C_1, C_2, ... C_i representa caracteres, intervalos de caracteres (c_1-c_2) ou classes de caracteres
[^...]	Complemento da classe de caracter
[...&&...]	Intersecção entre duas classes de caracter
Classes de Caractere Pré-definidas	
.	Qualquer caractere, exceto fim de linha
\d	Um dígito [0-9]
\D	Um "não" dígito [^0-9]
\s	Caractere espaço em branco [\t\n\r\f\x0B]
\S	Caractere "não" espaço em branco
\w	Um caractere de texto [a-zA-Z0-9_]
\W	Um "não" caractere de texto
\p{nome}	Uma classe de caractere nomeada
\P{nome}	O complemento de uma classe de caractere nomeada
Correspondências de Fronteira	
\b	A fronteira de uma palavra
\B	A fronteira de uma "não" palavra
\A	Início da entrada
\z	Fim da entrada
\Z	Fim da entrada, exceto fim de linha
\G	Final da última localização
Quantificadores	
X?	X é ocional
X*	X, 0 ou mais vezes
X+	X, 1 ou mais vezes
X{n} X{n,} X{n,m}	X *n* vezes, a menos *n* vezes, entre *n* e *m* vezes

Tabela 2 - Sintaxe de algumas expressões regulares em Java™

Categoria	Expressão Regular
Grupo 1 – contas devedoras	
Conta devedora zerada	.*total a pagar 0,00.*
Conta devedora 1 dígito	.*total a pagar \\d{1},\\d{2}[^c].*
Conta devedora 2 dígitos	.*total a pagar \\d{2},\\d{2}[^c].*
Conta devedora 3 dígitos	.*total a pagar \\d{3},\\d{2}[^c].*
Conta devedora 4 dígitos ou mais	.*total a pagar.*\\d{1,}\\.\\d{1},\\d{2}[^c].*
Grupo 2 - contas credoras	
Conta credora 1 dígito	.*total a pagar \\d{1},\\d{2}cr
Conta credora 2 dígitos	.*total a pagar \\d{2},\\d{2}cr
Conta credora 3 dígitos	.*total a pagar \\d{3},\\d{2}cr
Conta credora 4 dígitos ou mais	.*total a pagar .*\\d{1,}\\.\\d{4},\\d{2}cr
Grupo 3 - contas totalizadoras	
Conta Totalizadora (1 conta)	.*30 dias.*
Conta Totalizadora (2 contas)	.*mais de 60 dias.*
Conta Totalizadora (3 contas)	.*mais de 90 dias.*
Conta Totalizadora (4 ou mais contas)	.*Não perca sua linha.*
Clientes última conta credora	.*Banco do Brasil.*
Conta Simplificada	.*Simplificada.*
Grupo 4 – produtos e serviços	
Linha *a*	.*a.*
Linha *b*	.*b.*
Linha *c*	.*c.*
Habilitação de linha	.*Habilit.*
Doações	.*Doação.*|.*LBV.*|.*Lar de Maria.*|.*Boldrini.*|.*UNESCO.*|.*Teleton.*

Tabela 3 – Exemplos de categorias e expressões regulares parametrizadas para o primeiro exemplo

O exemplo foi conduzido em um computador do tipo PC, com um processador de 2.26 GHz, 1 GB de memória RAM, sistema operacional Windows XP e máquina virtual Java™ versão 5.0. O arquivo de impressão utilizado possui 376 MB de tamanho e mais de 57.000 páginas. O dicionário de expressões regulares soma um total de 105 expressões, as quais contemplam os itens da lista de verificação de duas equipes de analistas: serviços periódicos/ (os quais desempenham atividades distintas no tocante a análise do spool de impressão).

O algoritmo apresentado na Figura 12 foi aplicado ao *spool* de impressão durante 4,5 horas e interrompido de forma abrupta por questões práticas. A Tabela 4 apresenta as estatísticas gerais do primeiro exemplo.

A primeira coluna da tabela, "Qtd. De categorias", agrupa a quantidade de categorias que o classificador associou a uma página, independentemente das categorias. Por exemplo, suponha que uma página satisfaça a uma busca pelos padrões "*.*Simplificada.**" e "*.*Habilit.**"; e, na tabela, há uma linha que quantifique 2 categorias com as respectivas estatísticas. Essas 2 categorias referem-se a qualquer combinação de duas categorias. A coluna "Qtd. De Páginas" sumariza a quantidade de páginas por grupo "Qtd. de categorias". A unidade de medida do tempo apresentado está em segundos.

Qtd. de categorias	Qtd. de Páginas	Tempo Total (s)	Menor tempo (s)	Maior tempo (s)	Tempo médio (s)
14	2	2,64	1,31	1,33	1,32
13	6	6,63	0,70	1,64	1,10
12	4	3,83	0,58	1,33	0,96
11	20	25,80	0,63	2,19	1,29
10	47	51,81	0,52	1,70	1,10
9	73	88,02	0,53	1,89	1,21
8	195	240,96	0,48	2,00	1,24
7	337	412,57	0,44	2,20	1,22
6	568	700,16	0,41	2,02	1,23
5	768	942,34	0,36	2,73	1,23
4	1.161	1.429,61	0,39	2,91	1,23
3	3.703	4.360,46	0,38	3,88	1,18
2	3.011	3.647,89	0,30	4,16	1,21
1	630	1.083,37	0,25	4,55	1,72
0	5.306	2.927,91	0,00	4,69	0,55
Total geral	15.831	15.923,98	0,00	4,69	1,01

Tabela 4 - Estatísticas gerais do primeiro exemplo

4.1.1 Análise Crítica do Primeiro Exemplo

Foi verificado na prática que, a construção de expressões regulares é relativamente complexa, necessitando de um período inicial de aprendizagem tanto na sintaxe das expressões quanto em experimentos de tentativa e erro para obtenção de experiência.

Apesar de ser um protótipo, o desempenho do classificador pode ser considerado regular, classificando um total de 15.831 páginas, com um tempo médio de 1,01

segundo por página. Considerando uma extrapolação para um total de 57.000 páginas, o tempo necessário estimado para classificação do arquivo como um todo ficaria em torno de 15,9 horas, algo que deixa a desejar considerando o prazo de análise dos analistas. Ao conduzir a mesma experiência, utilizando o mesmo arquivo de impressão, o mesmo classificador, e reduzindo-se a quantidade de expressões regulares (foi considerado as expressões de apenas uma equipe de analistas), o desempenho melhorou de 1,01 segundo para 0,52 segundos médios. Desta forma depreende-se que, a utilização de um computador com melhor capacidade de processamento e um projeto de classificador mais rebuscado pode trazer ganhos significativos ao processo.

4.2 Segundo Exemplo

O principal objetivo do segundo exemplo foi verificar a possibilidade de se obter melhores resultados, por meio da melhoria na representação das páginas do *spool* de impressão.

Para condução do segundo exemplo, cada página do spool de impressão foi transformada numa página de texto simples, considerando a ordem de apresentação dos termos. Utilizando-se dos conceitos e técnicas apresentados no item 3.6.1, "Técnica para Identificação de Sentenças Comparativas", cada palavra-chave do texto foi substituída por uma marca (*token*) – na verdade, apenas as palavras[5] correspondentes a um valor monetário (i.e. corresponde a determinado padrão).

Por exemplo, toda palavra correspondente a expressão regular *\\d{1},\\d{2}* foi convertida na marca *numero1.2*. Na prática, a cadeia de caracteres "*total a pagar 132,45*" é transformada em "*total a pagar numero3.2*". A Tabela 1 lista as expressões regulares e respectivas marcas utilizadas para transformação do *spool* de impressão. A Figura 13 ilustra o processo de transformação.

[5] No sentido desta obra, uma palavra é uma cadeia de caracteres sem espaços em branco

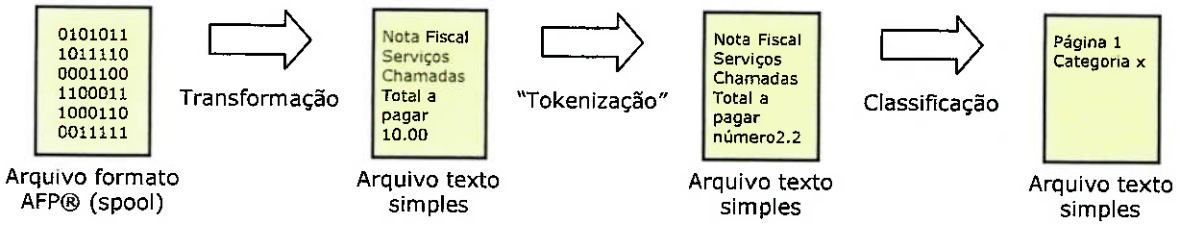


Figura 13 - Processo de transformação do *spool* de impressão

Importante ressaltar que, na condução do exemplo, a aplicação de uma expressão regular implicou na exclusão das demais, isto é, as expressões são mutuamente excludentes.

Expressão Regular	Marca	Exemplo
0,00	0,00	0,00
\\d{1},\\d{2}	numero1.2	0,00 ~ 9,99
\\d{2},\\d{2}	numero2.2	10,00 ~ 99,99
\\d{3},\\d{2}	numero3.2	100,00 ~ 999,99
.*\\d{1,}\\.\\d{3},\\d{2}	numero4+.2	1.000,00 ~ *.999,99
\\d{1},\\d{2}cr	numero1.2cr	0,00 ~ 9,99cr
\\d{2},\\d{2}cr	numero2.2cr	10,00 ~ 99,99cr
\\d{3},\\d{2}cr	numero3.2cr	100,00 ~ 999,99cr
.*\\d{1,}\\.\\d{3},\\d{2}cr	numero4+.2cr	1.000,00 ~ *.999,99cr

Tabela 5 - Expressões e marcas utilizadas para transformação

Com base na lista de verificação dos analistas de faturamento e, alinhado às marcas definidas, foi criado um dicionário de categorias e seqüências-chave para utilização no exemplo. A Tabela 6 apresenta algumas das 116 categorias e seqüências-chave parametrizadas para o exemplo.

Categoria	Seqüência-chave
Grupo 1 – contas devedoras	
Conta devedora zerada	total a pagar 0,00
Conta devedora 1 dígito	total a pagar numero1.2
Conta devedora 2 dígitos	total a pagar numero2.2
Conta devedora 3 dígitos	total a pagar numero3.2
Conta devedora 4 dígitos ou mais	total a pagar numero4+.2
Grupo 2 - contas credoras	
Conta credora 1 dígito	total a pagar numero1.2cr
Conta credora 2 dígitos	total a pagar numero2.2cr
Conta credora 3 dígitos	total a pagar numero3.2cr
Conta credora 4 dígitos ou mais	total a pagar numero4+.2cr

Categoria	Seqüência-chave
Grupo 3 - contas totalizadoras	
Conta Totalizadora (1 conta)	30 dias
Conta Totalizadora (2 contas)	mais de 60 dias
Conta Totalizadora (3 contas)	mais de 90 dias
Conta Totalizadora (4 ou mais contas)	Não perca sua linha
Clientes última conta credora	Banco do Brasil
Conta Simplificada	Simplificada
Grupo 4 – produtos e serviços	
Linha *a*	a
Linha *b*	b
Linha *c*	c
Habilitação de linha	habilit
Doações	Doação
LBV	LBV
Lar de maria	Lar de maria
Boldrini	Boldrini

Tabela 6 - Categorias e seqüências-chave para classificação

Formalmente, o arquivo transformado D é um arquivo de texto simples composto por páginas de texto, de forma que $D = \{d_1, \ldots, d_{|D|}\}$; onde cada página d é composta por linhas de texto $L = \{l_1, \ldots, l_{|L|}\} \subset d$ e, cada linha é composta de termos de forma que $l_i = \{w_1, \ldots, w_{|l|}\} \subset d$. Seja C um conjunto de pares "categoria/seqüência-chave" de forma que $C = \{\langle c_1, s_1 \rangle, \ldots, \langle c_{|C|}, s_{|C|} \rangle\}$. Cada página foi classificada por k categorias de forma que $0 \leq k \leq |C|$. A Figura 12 apresenta o algoritmo utilizado para classificação das páginas.

```
for (each p in D)                                    ①
{
        for (each c in C)                            ②
        {
                if (p contains c.sequence) then      ③
                        assign c.category to p;
                next c;
        }
        next p;
}
```

Figura 14 – Algoritmo aplicado no segundo exemplo para classificação das páginas

O algoritmo da Figura 14 deve ser interpretado da seguinte forma:

a) O primeiro laço (①) itera sobre cada página da coleção de documentos;

b) O segundo laço (②) itera sobre cada par categoria-seqüência-chave da coleção de categorias, para cada página da coleção;

c) O teste condicional (③) avalia se a página contém a seqüência-chave daquela iteração. Em caso afirmativo a categoria é atribuída à página, caso contrário avalia-se a próxima categoria.

O algoritmo supracitado foi aplicado nas mesmas condições do primeiro exemplo, isto é, mesmo hardware, sistema operacional, máquina virtual Java™ e *spool* de impressão. Após 22,5 minutos de execução, todas as páginas do *spool* de impressão foram classificadas. A Tabela 7 apresenta as estatísticas gerais do segundo exemplo.

A primeira coluna da tabela, "Qtd. De categorias", agrupa a quantidade de categorias que o classificador associou a uma página, independentemente das categorias. Por exemplo, suponha que uma página satisfaça a uma busca pelas seqüências "*total a pagar numero1.2*" e "*Habilit*"; e, na tabela, há uma linha que quantifique 2 categorias com as respectivas estatísticas. Essas 2 categorias referem-se a qualquer combinação de duas categorias. A coluna "Qtd. De Páginas" sumariza a quantidade de páginas por grupo "Qtd. de categorias". A unidade de medida do tempo apresentado está em milissegundos (milésima parte de um segundo).

Qtd. de categorias	Qtd. de Páginas	Tempo Total (ms)	Menor tempo (ms)	Maior tempo (ms)	Tempo médio (ms)
14	1	15	15	15	15,0
13	6	156	15	32	26,0
12	7	201	15	62	28,7
11	27	714	0	94	26,4
10	79	1.741	0	47	22,0
9	198	4.659	0	63	23,5
8	410	10.115	0	172	24,7
7	854	20.734	0	156	24,3
6	1.748	43.915	0	328	25,1
5	3.789	80.325	0	313	21,2
4	11.277	192.074	0	656	17,0
3	6.089	121.779	0	469	20,0
2	11.354	183.373	0	547	16,2
1	2.343	79.312	0	546	33,9
0	19.020	203.020	0	953	10,7

Qtd. de categorias	Qtd. de Páginas	Tempo Total (ms)	Menor tempo (ms)	Maior tempo (ms)	Tempo médio (ms)
Total geral	57.202	942.133	0	953	16,5

Tabela 7 - Estatísticas gerais do segundo exemplo

4.2.1 Análise Crítica do Segundo Exemplo

Os resultados apresentados na Tabela 7 mostram que a utilização de marcas ao invés de expressões, e, modificação do classificador para localizar seqüências-chave num corpo de texto trouxe ganhos significativos de desempenho. Todo o *spool* de impressão foi classificado em apenas 22,5 minutos, com uma média de 16,5 milissegundos por página. Do ponto de vista prático, a utilização de seqüências-chave ao invés de expressões regulares mostrou-se mais simples e fácil de usar, tanto para um usuário final como para o desenvolvimento de um classificador.

5 Conclusões e Trabalhos Futuros

5.1 Conclusões

O primeiro exemplo considerou a utilização de expressões regulares para classificação de cada página do *spool* de impressão. Na prática, a utilização de expressões regulares demonstrou-se bastante complexa para utilização de usuários finais - no caso do problema proposto, analistas de faturamento. Inicialmente foi necessário dedicar algum tempo para: (1) familiarização com a sintaxe das expressões regulares; (2) obtenção de experiência na construção de expressões por meio de "tentativa-e-erro"; (3) ajustes no dicionário de expressões regulares do classificador.

O primeiro classificador classificou um total de 15.831 páginas, com um tempo médio de 1,01 segundo por página. Extrapolando este desempenho para um total de 57.000 páginas, o tempo necessário estimado para classificação do arquivo como um todo ficaria em torno de 15,9 horas, um desempenho regular considerando o processo de análise do *spool* de impressão. Poder-se-ia sugerir a classificação do *spool* de impressão fora do expediente normal, ou a utilização de um processador mais veloz para obter melhores resultados.

Já o segundo exemplo considerou a transformação do *spool* de impressão, onde (1) determinadas palavras-chave foram substituídas por marcas, com o objetivo de se obter seqüências-chave, e; (2) modificação do classificador para localizar seqüências-chave. Essas mudanças trouxeram ganhos significativos de desempenho, de forma que todo o *spool* de impressão foi classificado em apenas 22,5 minutos, com uma média de 16,5 milissegundos por página.

Do ponto de vista prático, a utilização de seqüências-chave ao invés de expressões regulares mostrou-se mais simples e fácil de usar, tanto para um usuário final como para o desenvolvimento de um classificador, mostrando, em suma, que uma revisão dos conceitos e do projeto do software pode evitar investimentos desnecessários.

Depreende-se também que, a economia de tempo em pesquisas resulta em um ganho intangível aos analistas de faturamento, os quais podem dedicar esse tempo em atividades mais nobres, como a análise das contas telefônicas, por exemplo. Não obstante, a companhia tem um incremento de qualidade no processo, o qual pode ser traduzido num maior volume de contas analisadas.

5.2 Trabalhos Futuros

A presente obra considerou tão somente o desenvolvimento de um protótipo de ferramenta para solucionar um problema da operação de faturamento de uma operadora de telecomunicações do grupo Telefônica: a localização de contas telefônicas no *spool* de impressão para análise dos analistas de faturamento.

Vale lembrar que, conforme descrito no capítulo 2, a atividade de análise de contas telefônicas em formato digital envolve: (1) localizar os itens da lista de verificação no *spool* de impressão e (2) análise das contas telefônicas de acordo com as instruções de trabalho, padrões de qualidade, normas internas etc.

Há um campo bastante interessante para pesquisas futuras, o qual consiste em automatizar/semi-automatizar algumas análises realizadas pelos analistas de faturamento, como por exemplo, verificar se os subtotais dos grupos de itens na fatura correspondem à soma dos itens individualmente, ou se a tarifa aplicada a determinado produto/serviço está correta, ou se todos os itens valorados na Nota Fiscal Fatura de Serviços de Telecomunicações estão numerados, etc.

REFERÊNCIAS

APTÉ, C.; DAMERAU F.; WEISS, S. M. Automated Learning of Decision Rules for Text Categorization. ACM Transactions on Information Systems, v. 12, n. 3, p. 233-251, 1994

BRILL, E. A Simple Rule-base Part of Speech Tagger. ANL, 1992

HORSTMANN, C., CORNELL, G. **Core Java 2 – Fundamentals**. 7th ed. Santa Clara-CA: Sun Microsystems Press, 2005, 2 v. in 1.

JINDAL, N.; BING, L. Identifying Comparative Sentences in Text Documents. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006, Seattle, Washington. New York: ACM, 2006. p. 244-251

LEWIS, D. D. Feature Selection and Feature Extraction for Text Categorization. In: Proceedings of the workshop on Speech and Natural Language, 1992, New York. San Mateo, CA: Morgan Kaufmann, 1992. p. 212-217

MITCHELL, T.M. **Machine Learning**. New York: McGraw Hill, 1996

SALTON, G.; BUCKLEY C. Term-weighting Approaches in Automatic Text Retrieval. Information Processing & Management, v. 24, n. 5, p. 513-523, 1988

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

SEBASTIANI, F. Machine Learning in Automated Text Categorization. ACM Computing Surveys, v. 34, n. 1, p. 1-47, 2002

SPANGLER, S.; KREULEN J. Interactive Methods for Taxonomy Editing and Validation. Proceedings of the Eleventh International Conference on Information and Knowledge Management, 2002, McLean, Virginia. New York: ACM, 2002. p. 665-668

SPANGLER, S.; KREULEN J. LESSER, J. MindMap: Utilizing Multiple Taxonomies and Visualization to Understand a Document Collection. Proceedings of the 35th Annual Hawaii International Conference on System Sciences, 2002. IEEE.

TELEFÔNICA BRASIL. Apresenta descrição sobre o grupo Telefônica no Brasil. Disponível em < http://www.telefonica.net.br/sobre/gbrasil.htm>.
Acesso em: 02 dez. 2006.

TELEFÔNICA S.A. Apresenta descrição sobre o grupo Telefônica. Disponível em: <http://www.telefonica.es/acercadetelefonica/por/1descripcion/actividad.shtml>.
Acesso em: 02 dez. 2006.

WITTEN, I. H.; FRANK E. **Data Mining: Practical Machine Learning Tools and Techniques**. 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

REFERÊNCIAS COMPLEMENTARES

ALIPPI, C.; PESSINA, F.; ROVERI, M. An Adaptive System for Automatic Invoice-Documents Classification. IEEE International Conference on Image Processing, 2005, v. 2, p. 526-529, Washington: IEEE, 2005

DEBRY, R. K.; PLATTE, R. G. Advanced Function Printing: A Tutorial. IBM Systems Journal, v. 27, n. 2, p. 219-233, 1998

DHILLON, I. et al. Visualizing Class Structures of Multi-Dimensional Data. Proceedings of 30th Conference on Interface, Computer Science and Statistics, 1998

DORAN, C.; EGEDI, D.; HOCKEY, B. A.; SRINIVAS, B., ZAIDEL, M. XTAG System – A Wide Coverage Grammar for English. COLING'94, 1994

GAO, X.; MURUGESAN, S. Extraction of Keyterms by Simple Text Mining for Business Information Retrieval. Proceedings of the 2005 IEEE International Conference on e-Business Engineering, v. 18, n. 21, p. 332-339, Washington: IEEE, 2005

IBM Printing Systems Division. **Mixed Object Document Content Architecture Reference**. 8th ed. Boulder-CO: 2006

JAILLET, S.; TEISSEIRE, M.; CHAUCHE, J. PRINCE, V. Classification of Documents by Content. Proceedings of the Second IEEE International Conference on Cognitive Informatics, v. 18, n. 20, p. 214-222, 2003

LAM, W.; LOW, K. Automatic Document Classification Based on Probabilistic Reasoning: Model and Performance Analysis. 1997 IEEE International Conference on Systems, Man, and Cybernetics, v. 3, p. 2719-2723, Washington: IEEE, 1997

LARKEY, L. S.; CROFT, W. B. Combining Classifiers in Text Categorization. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, New York : ACM, 1996

MITCHELL, T.M. **Machine Learning**. New York: McGraw Hill, 1996

MOULINIER, I.; GANASCIA, J.; Applying an Existing Machine Learning Algorithm to Text Categorization. Lecture Notes In Computer Science, v. 1040, p. 343-354 London : Springer-Verlag, 1996

MÜLLER, A.; DÖRRE, J.; GERSTL, P.; SEIFFERT, R. The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection. Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences, Washington : IEEE, 1999

OLSEN, K. A.; KORFHAGE, R. R.; SOCHATS, K. M.; SPRING, M. B.; WILLIANS, J.G. Visualization of a Document Collection: The VIBE System. Information Processing & Management, v. 29, n. 1, p. 69-81, 1993

SAKURAI, S.; UENO, K. Analysis of Daily Business Reports Based on Sequencial Text Mining Method. 2004 IEEE International Conference on Systems, Man and Cybernetics, v. 4, p. 3279-3284, Washington: IEEE, 2004

STREHL, A. **Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining**. 232 p. 2002. Dissertação (Phd) - University of Texas at Austin

ZAÏANE, O. R.; ANTONIE, M. Classifying Text Documents by Associating Terms with Text Categories. Proceedings of the 13th Australasian database conference, v. 5, p. 215-222, Melborne : Australian Computer Society, 2002

ANEXO A – Modelo de Check-list Aplicado na Atividade de Análise do Spool de Impressão

Analista Responsável: *Nome do Analista*

Dados sobre o arquivo
Tamanho: *tamanho em MB*
Qtd. Páginas: *número de páginas*
Data de disponibilização: *dd/mm/aaaa*

2 - Valores Mensais	Check
Assinatura de Plano Básico ANATEL	
Assinatura de Plano Alternativo ANATEL	
Assinatura de Planos de minutos livres	
- Plano de minutos A	
- Plano de minutos B	
- Plano de minutos C	
Linha AICE - Acesso Individual Classe Especial	

3 - Outros Serviços	Check
Promoção X	
Bloqueador Celular	
Bloqueador de Interurbano	
Bônus de Assinatura	
Caixa Postal Multipla	
Chamada a três	
Consulta e Conferência	
Consulta e Transferência	
Correção de instalação	
Crédito por interrupção do STFC	
Discagem abreviada	
Internet Banda Larga	
Aluguel de Modem	
Substituição do número telefônico	
Transferência de chamada	
Transferência linha ocupada	

4 - Tráfego Próprio	Check
Chamadas Locais a Cobrar - Fixo / SME / SMP / SMC	
Intra - Fixo / SME / SMP / SMC	
Inter - Fixo / SME / SMP / SMC	

5 - Tráfego de Outras Operadoras	Check
Chamada 0300	
Intra - Fixo / SME / SMP / SMC	
Inter - Fixo / SME / SMP / SMC	