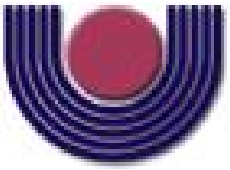


Unioeste - Universidade Estadual do Oeste do Paraná
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
Colegiado de Ciência da Computação
Curso de Bacharelado em Ciência da Computação

Uma aplicação de Mineração de Texto em Faturas de Telefonia Móvel Corporativa

Guilherme Felipe Zabot

CASCAVEL
2016

Guilherme Felipe Zabot

Uma aplicação de Mineração de Texto em Faturas de Telefonia Móvel Corporativa

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel

Orientador: Prof. Dr. Clodis Boscarioli

CASCAVEL
2016

Guilherme Felipe Zabot

Uma aplicação de Mineração de Texto em Faturas de Telefonia Móvel Corporativa

Monografia apresentada como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Prof. Dr. Clodis Boscarioli (Orientador)
Colegiado de Ciência da Computação,
UNIOESTE

Gustavo Rezende Krüger (Coorientador)
Orb it Sistemas, Cascavel-PR

Prof. Dr. Marcio Seiji Oyamada
Colegiado de Ciência da Computação,
UNIOESTE

Cascavel, 14 de fevereiro de 2017

DEDICATÓRIA

Dedico este trabalho à meus pais e familiares por sempre me incentivarem nesta árdua caminha.

AGRADECIMENTOS

Aos meus pais, Lari José Zabot e Tânia Mara Coppi Zabot, que sempre me apoiaram nas minhas decisões e sempre que precisei estiveram ao meu lado, oferecendo tudo o que podiam para alcançar meus objetivos.

À minha namorada, Angélica Dettoni Modzinski, que em todos estes anos de graduação sempre esteve ao meu lado me ajudando e dando apoio nas horas mais difíceis, ajudando-me também nas minhas escolhas e decisões, e mesmo longe, nunca me deixou sozinho.

Agradeço aos professores do curso de Ciência da Computação, que sempre de disponibilizaram a ajudar na minha formação, em especial ao meu orientador professor Clodis Boscarioli, que caminhou ao meu lado durante toda a graduação, sendo orientador dos meus projetos de iniciação científica e também orientador deste trabalho.

Agradeço também a meu coorientador Gustavo Rezende Krüger, pela oportunidade de estágio proporcionada durante meu ultimo ano na graduação, e pela ajuda na elaboração deste trabalho.

Agradeço a todos os meus amigos e colegas, que de alguma forma me apoiaram e contribuíram para a conclusão do curso.

Quero ainda, agradecer aos colegas membros do Grupo PETComp (Programa de Educação Tutorial em Ciência da Computação), e todos que colaboraram direta ou indiretamente com mais este passo em minha vida.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

EI	Extração de Informação
HTML	*HypertText Markup Language*
KDD	*Knowledge-Discovery in Databases*
KDT	*Knowledge Discovery from Text*
MD	Mineração de Dados
MT	Mineração de Texto
PDF	*Portable Document Format*
PLN	Processamento de Linguagem Natural
RI	Recuperação de Informação
SGBD	Sistema Gerenciador de Banco de Dados
TXT	Arquivos em formato textual

Sumário

Resumo

A Extração de Informação (EI) é uma coleção de métodos e técnicas que têm como objetivo extrair, de fontes semiestruturadas ou não estruturadas, informações relevantes. Um sistema de EI é capaz de extrair de fontes textuais apenas informações que seja de interesse dos usuários do sistema, visto que as partes insignificantes não são extraídas. Neste trabalho são apresentados e discutidos os principais conceitos, técnicas e abordagens computacionais da área. Também é proposta uma aplicação supervisionada de EI utilizando a abordagem de aprendizado de máquina em Faturas de Telefonia Móvel Corporativa, que apresentam em seu conteúdo uma organização hierárquica. A aplicação é composta por módulos de identificação do problema, pré-processamento, extração de padrões e pós processamento. Cada um dos módulos é discutido, assim como as técnicas utilizadas para implementá-los. O sistema de extração de informação é avaliado experimentalmente visando mensurar sua precisão. Os experimentos realizados mostram resultados consideráveis, que podem ser tomados como base para o desenvolvimento de outros sistemas capazes de extrair informações em documentos textuais.

Palavras-chave: Extração de informação, Mineração de Textos, Classificadores.

Capítulo 1

Introdução

A quantidade de dados armazenados em repositórios digitais vem crescendo a taxas elevadas. Muitas dessas informações encontram-se em formatos não estruturados ou semiestruturados, em documentos textuais como arquivos em formato PDF ou em páginas Web de formato *HypertText Markup Language* (HTML), desenvolvidos para possibilitar a compreensão por seres humanos, não sendo adequados à análise de informação. Neste contexto, é cada vez maior a necessidade de métodos para a extração de informação nesses documentos.

Emerge aí a área de Mineração de Texto (*Text Mining* - MT), também conhecida como (*Knowledge Discovery from Texts* - KDT), cujos processos foram descritos por [Feldman e Dagan 1995], e faz uso de técnicas de análise e extração de dados a partir de coleções de documentos textuais.

A Mineração de Textos é um processo análogo à Mineração de Dados (*Data Mining* - MD), que é parte do processo de descoberta de conhecimento em base de dados (*Knowledge Discovery from Databases* - KDD), onde no primeiro caso as fontes de dados são textos não estruturados ou semiestruturados e no segundo caso as fontes de dados são estruturadas [Feldman e Dagan 1995].

A Mineração de Texto envolve a aplicação de algoritmos que processam textos e identificam informações úteis, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, dado que a informação contida nesses textos não pode ser obtida de forma direta devido ao seu formato.

Segundo [Silva 2004], EI é um processo capaz de extrair conceitos, estatísticas e palavras relevantes de documentos textuais, a fim de transformá-los em um formato estruturado, como uma estrutura tabular, de forma a reduzir a informação não estruturada contida no documento.

Os sistemas baseados em EI não têm por objetivo compreender os textos processados, e sim realizar a extração dos dados relevantes ao usuário. Esses sistemas devem modelar uma função que receba um documento de entrada e retorne como saída um formulário com seus respectivos campos preenchidos.

De acordo com [Appelt e Israel 1999], a construção de um sistema de EI pode ser realizada por meio de duas abordagens: a engenharia do conhecimento e o aprendizado de máquina. Na primeira, uma pessoa familiarizada com o formalismo usado para codificar as regras de extração e que possua conhecimento sobre os sistemas de EI deve, sozinho ou com a ajuda de um especialista no domínio, codificar manualmente as regras de extração. Na segunda abordagem, utiliza-se um algoritmo de aprendizado de máquina para aprender automaticamente essas regras a partir de um conjunto de documentos.

Para [Silva 2004], a escolha de uma abordagem para a construção de um sistema de EI não é trivial, e o formato de texto é importante para a definição das técnicas que devem ser utilizadas no processo de EI, devido à aplicabilidade de cada técnica a um determinado tipo de texto.

Os problemas da Extração de Informação podem ser resolvidos por classificadores, que podem ser vistos como uma caixa preta capaz de classificar uma entrada de texto, representado por um vetor de características, em uma categoria ou classe predefinida. Assim, o algoritmo recebe do documento de entrada um fragmento de texto, e determina o grau de confiança que esse fragmento preenche corretamente cada um dos campos do formulário de saída.

Os benefícios da MT podem se estender a qualquer domínio que utilize algum tipo de texto, como é o caso da Gestão de Custos em Telecomunicações. De acordo com [Cavalcante 2009], o profissional da área de Gestão de Custos em Telecomunicações tem o objetivo de compreender, gerenciar e otimizar a utilização dos serviços de telecomunicações dentro de um ambiente corporativo. Para tanto, faz-se necessária a interpretação das faturas de telecomunicações fornecidas pelas operadoras prestadoras de serviços.

Por se tratarem de documentos extensos e textuais, a leitura, compreensão e síntese dos dados das faturas tornam-se manualmente inviáveis e quando viáveis, a chance do erro humano aumenta consideravelmente. Pretendendo automatizar o processo de estruturação das informações presentes nesses documentos, este trabalho visa auxiliar na tomada de decisão na área de Gestão de Custos de Telecomunicações.

Na literatura, encontram-se vários trabalhos relacionados à extração e recuperação de informações de resumos, de referências bibliográficas ou artigos científicos, bem como trabalhos referentes à área da Bioinformática, no sentido da análise da informação na área de estudos da Biologia. Alguns dos trabalhos relacionados à área de Extração de Informação e Recuperação de Informação, aplicadas a esses contextos, e utilizadas como base para esse trabalho, podem ser visualizados na Seção 2.3. Não foram encontrados trabalhos referentes a EI relacionados à área de telecomunicações, mais precisamente a EI em faturas de telefonia móvel corporativa.

1.1 Objetivos

Desenvolver um sistema que realize a Extração de Informações de faturas de telefonia móvel corporativa, que se encontram em formato PDF. Os objetivos específicos são:

- Organizar e estruturar os dados das faturas em um formato tabular.
- Apresentar um sistema baseado em Extração de Informação e uma abordagem de Classificação Hierárquica.

1.2 Estrutura do Trabalho

O presente trabalho encontra-se dividido em seis capítulos.

O Capítulo 1 Introdução, trouxe o contexto, a motivação e objetivos do trabalho.

O Capítulo 2, Mineração de Textos, apresenta inicialmente conceitos sobre a Mineração de Texto e seus processos, trabalhos relacionados à área de extração de informação e classificação hierárquica, seguido do embasamento teórico que permite compreender os aspectos e elementos de MT necessários à construção do sistema proposto.

No Capítulo 3 Modelo Proposto, são apresentados os processos necessários para a construção do sistema de Extração de Informação.

O Capítulo 4, Construção do Modelo, apresenta os passos realizados na construção do modelo proposto para o sistema, bem como a definição dos processos realizados e o formulário de saída esperado.

O Capítulo 5, Experimentos e Resultados, descreve como foram elaborados os experimentos e os dados utilizados, seguido de uma análise dos resultados com a aplicação da técnica de

aprendizado de máquina, comparando os resultados obtidos sob diferentes conjuntos de treinamento.

O Capítulo 6, Considerações Finais, contempla as contribuições deste trabalho, bem como as possíveis aplicações para trabalhos futuros, visando melhorar o sistema desenvolvido.

Capítulo 2

Mineração de Textos

Este capítulo apresenta uma introdução à área de Mineração de Textos, com foco principal na Extração de Informação e Classificação de Informações. As Seções 2.1 e 2.2 trazem os conceitos básicos da área de MT e da área de EI, respectivamente. A Seção 2.3 apresenta os trabalhos relacionados, e a Seção 2.4 apresenta as considerações finais do capítulo.

2.1 Conceitos Básicos

A MT se refere ao processo não trivial de extração de conhecimento útil em uma coleção de documentos textuais, atuando de forma semelhante à Mineração de Dados. Pode ser definida como um conjunto de técnicas e processos utilizados para descobrir padrões interessantes em um conjunto de documentos textuais não estruturados [Feldman e Hirsh 1997], possuindo métodos inteligentes e ferramentas automáticas para auxiliar pessoas na análise de grandes volumes de textos, a fim de minerar o conhecimento útil.

Apesar de a MT, de maneira análoga a MD, buscar extrair informação útil de bases de dados pela identificação e exploração de padrões interessantes, deve-se destacar a principal diferença entre ambas. A MT atua a partir de bases textuais que possuem pouca ou nenhuma estrutura de dados e que, eventualmente, possuem ambiguidade. Já a MD, realiza a obtenção de conhecimento em bases de dados bem estruturadas [Dorre, Gerstl e Seiffert 1999].

A MT é considerada uma área multidisciplinar, pois utiliza técnicas das áreas de EI, Processamento de Linguagem Natural (PLN) e Recuperação de Informação, juntamente com algoritmos de Aprendizado de Máquina e Estatística. As principais contribuições desta área estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quan-

titativa de grandes volumes de textos, e à melhor compreensão dos conteúdos disponíveis em documentos, que podem estar em diferentes formatos, como e-mails, páginas Web e arquivos em formato PDF.

Tendo em vista que a área de MT têm focado em problemas de representação de textos, classificação, agrupamento, EI e busca ou modelagem de padrões, a seleção de atributos relevantes e a influência do conhecimento do domínio são de suma importância.

2.1.1 O Processo de Mineração de Textos

O processo de mineração de textos, segundo [Rezende 2003], pode ser dividido em cinco etapas, como pode ser observado na Figura 2.1, e abaixo explicadas.

Figura 2.1: Processo de mineração de textos
Fonte: [Rezende 2003]

- **Identificação do Problema:** De acordo com [Corrêa, Marcacini e Rezende 2012], a etapa da Identificação do Problema, é responsável basicamente por determinar o escopo do problema, onde dado objetivo da aplicação seleciona-se ou cria-se uma base de documentos e define-se o que se espera de sua análise.

  Esta é uma etapa fundamental para o decorrer da mineração, pois o conhecimento adquirido servirá como base para etapas seguintes do processo. Um especialista no domínio

elabora um estudo do domínio da aplicação, e a definição dos objetivos e metas a serem alcançados são identificados.

O especialista também deve identificar e selecionar a coleção de textos que será utilizada na extração do conhecimento.

- **Pré Processamento:** É uma etapa indispensável no processo de mineração de textos, pois a qualidade dos documentos pode influenciar no resultado final da extração do conhecimento. Visa estruturar os textos de maneira a torná-los processáveis por algoritmos de aprendizado de máquina.

  Após a realização da etapa de Identificação do Problema é feita a preparação dos textos, que consiste na padronização necessária para converter textos que estão em formatos como PDF entre outros, para um formato de texto plano e sem formatação. Realizada a padronização, elimina-se as *stopwords*, que são palavras consideras irrelevantes à representação da coleção.

  O próximo passo é a Extração de Termos, para isso pode-se utilizar algumas técnicas que simplificam suas formas de representação. Nesse cenário, podem ser utilizadas técnicas que reduzem o número de atributos ou constroem atributos mais representativos.

- **Extração de Padrões:** Na extração de padrões é escolhida a tarefa de mineração de dados que será empregada para que os objetivos da identificação do problema sejam atingidos. Ela é baseada no objetivo que foi definido no processo, e pode ser tanto preditiva como descritiva [Rezende 2003].

  As tarefas preditivas generalizam experiências passadas, como respostas conhecidas, em um modelo capaz de identificar a classe de um exemplo novo. Entre elas estão a classificação, que consiste em determinar a classe (ou categoria) a qual pertence um dado documento [Yang e Pedersen 1997].

  Em relação às descritivas, tem como objetivo identificar os comportamentos presentes no conjunto de dados, sendo que esses não apresentam rótulos específicos. Entre essas tarefas, pode-se destacar as Regras de associação, que tem como objetivo identificar associações entre registro de dados que estão ou deveriam estar relacionados, e agrupamento,

que busca segmentar populações heterogêneas em subgrupos ou segmentos homogêneos. Mais detalhes sobre essas tarefas são descritos em [Hand, Mannila e Smyth 2001].

- **Pós-Processamento:** Essa etapa visa avaliar, visualizar ou documentar para o usuário, as informações que foram descobertas na etapa de Extração de Padrões. O conhecimento extraído é validado por meio de avaliações, nas quais estar presente o especialista do domínio para verificar se o objetivo foi atingido.

- **Utilização do Conhecimento:** O conhecimento após ser validado e avaliado pelo especialista do domínio na etapa de Pós-Processamento é consolidado e pode ser incorporado a um Sistema Inteligente, ou ainda, utilizado diretamente para apoiar algum processo de tomada de decisão, conforme o objetivo definido na etapa de Identificação do Problema.

2.2 Tarefas da Mineração de Textos

Para realizar a MT há diversas tarefas como Recuperação de Informação (*Information Retrieval* - RI), Classificação/Categorização, Extração de Informação, Sumarização, Agrupamento, [Jackson e Moulinier 2002], dentre outras. Nessa seção uma descrição das tarefas pertinentes ao contexto dessa pesquisa será apresentada.

2.2.1 Extração de Informação

Há uma grande quantidade de dados em formato de textos eletrônicos, porém muitos desses não são tratados, pois nenhuma pessoa pode ler, compreender e sintetizar gigabytes de texto diariamente. Essas informações desperdiçadas têm estimulado pesquisadores a explorar novas estratégias para gerenciá-las. As estratégias mais comuns nesse meio são RI, filtragem de informações e EI.

Segundo [Kushmerick e Thomas 2003], a EI é uma forma de processar documentos, com o objetivo de popular uma base de dados com os valores extraídos automaticamente a partir de documentos.

A EI começa com a coleção de tais textos, transformando-os em informação e pronta para ser analisada. Os fragmentos de textos relevantes são isolados, a informação relevante é extraída destes fragmentos e então os pedaços são juntados coerentemente

[Zambenedetti e Oliveira 2002].

O objetivo da pesquisa em extração de informações é construir sistemas que encontrem e liguem informações relevantes e ignorem informações estranhas e irrelevantes [Cowie e Lehnert 1996]. Para [Riloff 1999], o propósito de sistemas de extração de informação é extrair informações de um domínio específico a partir de textos em linguagem natural.

Antes de apresentar os detalhes técnicos sobre sistemas para EI, é importante ressaltar que a escolha da técnica a ser usada na construção do sistema extrator depende da formatação do texto de entrada. Assim sendo, segue uma classificação de tipos de texto que irão guiar o resto da apresentação desta seção.

2.2.2 Tipos de Textos

O tipo de texto a partir do qual o sistema de EI irá realizar a extração das informações é importante para conseguir realizar a definição das técnicas que devem ser utilizadas no processo de extração. Na área de MT, os textos podem ser classificados em três tipos: textos estruturados, semiestruturados e não estruturados (livres), como visto em [Silva 2004]:

- **Texto Estruturado:** Um texto é considerado estruturado quando segue um formato predefinido e rígido. Por possuir regularidade em seu modelo, permite que suas informações sejam extraídas facilmente utilizando regras uniformes, baseadas, por exemplo, na ordem de seus elementos ou em delimitadores presentes no documento. Exemplos deste tipo de texto são páginas HTML de *e-commerce*, geradas automaticamente a partir de bancos de dados estruturados.

- **Texto Livre (Não Estruturado):** De maneira oposta aos textos estruturados, os textos livres não apresentam nenhuma regularidade em seu modelo. As informações nesses documentos aparecem como sentenças livres, escritas em uma linguagem natural. Por não possuir nenhuma estruturação em seus dados, a EI nesses tipos de textos não pode ser feita com base nas informações de formatação. Os sistemas de EI para textos livres geralmente utilizam técnicas de PLN, e as regras de extração são baseadas em padrões que envolvem as relações sintáticas entre as palavras, bem como as relações semânticas existentes entre elas [Soderland et al. 1995].

- **Texto Semiestruturado:** Tipo de texto presente em um ponto intermediário entre os textos estruturados e os textos livres. Não possuem formatação rígida, o que pode permitir a ocorrência de variações na ordem de seus dados. Além disso, em geral, não respeitam rigidamente a gramática da linguagem natural, e podem apresentar por exemplo palavras abreviadas. A Figura 2.2 possui um exemplo deste tipo de texto, contendo uma referência bibliográfica, que não segue um formato rígido, permitindo variações na ordem e na maneira como os as informações são apresentadas, bem como palavras abreviadas.

[Zhou 1996]ZHOU, E. Z. Object oriented programming c++ and power system simulation. *IEEE Transactions on Power Systems*, New York, v. 11, n. 1, p. 206–215, February 1996.

Figura 2.2: Exemplo de texto semiestruturado.

A seguir, visto que a explicação dos tipos de textos já foi abordada, uma breve apresentação sobre as técnicas utilizadas para a construção de sistemas de EI será dada.

2.2.3 Técnicas e Sistemas para Extração de Informação

Existem diferentes técnicas para a Extração de Informação. O tipo de texto de qual será feita a extração grande influência sobre a escolha da técnica a ser utilizada, que pode se basear apenas na estrutura do texto, quando existente. A literatura aponta duas técnicas principais para construção de sistemas para EI: os baseados em PLN e os baseados em *Wrappers*, como visto em [Silva 2004].

Em ambos os casos, são definidas regras ou padrões de extração para indicar os dados de interesse a serem extraídos. A definição de tais padrões pode ser feita manualmente, por algum especialista, ou com diferentes graus de automação. A seguir, detalhes sobre essas técnicas são apresentadas.

Sistemas para EI baseados em PLN

Os sistemas de EI baseados em PLN têm sido utilizados para diferentes domínios, no tratamento de documentos com pequeno ou nenhum grau de estruturação, contando com etapas de processamento de PLN em geral, e possuindo alguns módulos específicos para EI

[Cowie e Lehnert 1996]. O objetivo do uso dessas técnicas na EI é tentar compreender textos em alguma linguagem natural, a fim de encontrar os dados relevantes a serem extraídos.

Com base em [Appelt e Israel 1999], foram identificados seis módulos principais presentes em sistemas de EI baseados em PLN, que são: *tokenizador*, processador léxico, analisador sintático/semântico, extração de padrões, analisador do discurso, preenchimento de *templates*.

A fase de tokenização, ou separação de palavras, tem como objetivo separar os termos presentes no texto, de modo a possibilitar a identificação das sentenças. Geralmente, esta tarefa é realizada por meio do reconhecimento dos espaços em branco e de outros sinais, como pontuação.

Após a fase de separação das palavras, é realizada uma análise léxica e morfológica. Nesta etapa, a cada palavra é atribuída uma classe morfológica (substantivo, verbo, etc.) e outras características como reflexão de gênero, número e grau. Essa fase também é responsável pela identificação de nomes próprios e de outros itens que podem possuir uma estrutura interna, como data e hora.

A etapa de análise sintática e semântica é responsável por receber uma sequência de itens léxicos e tentar construir uma estrutura sintática para cada sentença do texto.A construção de regras de extração consiste na criação de um dicionário de regras de extração específico para o domínio que está sendo tratado, o que permite ao sistema extrair apenas os dados de interesse.

A etapa de análise do discurso tem como objetivo relacionar diferentes elementos do texto, não analisando frases isoladamente, e sim o texto como um todo, tentando identificar as influências de frases na interpretação de frases subsequentes.

O módulo de preenchimento de *templates* recebe os dados das etapas anteriores e preenche os *templates* definidos pela aplicação.

Os sistemas que utilizam PLN embutem em seu código algum conhecimento específico do domínio, o que o torna importante caso haja necessidade de alteração do domínio da aplicação, podendo-se alterar um domínio por completo com poucas mudanças no código do sistema. Em geral um sistema baseado em PLN dependerá do domínio a ser tratado, no caso deste trabalho, como os documentos encontram-se em formato semiestruturado e não seguem rigidamente a gramática da língua natural, optou-se por não utilizá-lo.

Sistemas Wrappers para EI

Os *Wrappers* exploram a regularidade apresentada por textos estruturados ou semiestruturados a fim de localizar os dados importantes. Em geral, um *Wrapper* tem como objetivo extrair as informações contidas em documentos e exportá-las como parte de uma estrutura de dados.

Normalmente um *Wrapper* é construído de maneira *ad-hoc*, não existindo uma arquitetura consensual para esse tipo de sistema. Antes de discutir como ocorre o processo de extração de informação executado pelos *Wrappers*, é importante apresentar as duas abordagens existentes para a implementação desses tipo de sistema: engenharia do conhecimento e aprendizado de máquina [Appelt e Israel 1999].

A abordagem baseada na engenharia do conhecimento é elaborada pela criação manual de regras de extração por um engenheiro do conhecimento, este é uma pessoa que já está familiarizada com o sistema de EI, e com o formalismo usado para codificar as suas regras de extração. Comumente o engenheiro do conhecimento tem acesso a um *corpus* do domínio de tamanho moderado, para que possa escrever e avaliar as regras de extração criadas.

Construir manualmente um *Wrapper*, significa escrever todo o código do programa. Esta abordagem geralmente requer grande quantidade de trabalho, visto que a construção é um processo iterativo. Neste processo o engenheiro do conhecimento escreve inicialmente um conjunto de regras, e avalia o resultado da extração, verificando se as regras criadas estão extraindo as informações corretamente ou não. Caso não esteja, ele modifica algumas regras, adiciona outras, e repete o processo, até que se atinja a taxa de acerto desejada.

A abordagem baseada em aprendizado de máquina é completamente diferente. Neste caso, não é necessária a existência de um engenheiro do conhecimento para codificar manualmente as regras do sistema de extração. É preciso apenas que exista alguém com conhecimento suficiente no domínio, e da tarefa de extração, para etiquetar um *corpus* de textos de treinamento e teste. Este etiquetamento consiste em determinar em cada texto, as informações que deverão ser extraídas pelo sistema. Um exemplo positivo de treinamento pode ser o documento completo que será analisado, ou mesmo apenas um pedaço dele, que caracteriza o elemento a ser extraído. De maneira análoga, um exemplo negativo, descreve um documento ou fragmento que não deverá ser coberto pelas regras de extração a serem aprendidas pelo sistema.

Dado que o *corpus* de treinamento foi criado de forma adequada, um algoritmo de aprendi-

zado de máquina é executado, resultando em algum tipo de conhecimento que o sistema de EI poderá utilizar para realizar a extração das informações. A forma de representação do conhecimento a ser aprendido pelo sistema depende da técnica de EI utilizada. O conhecimento pode ser representado na forma de uma expressão regular, um conjunto de regras de classificação, autômatos finitos, entre outros [Silva 2004].

Os *Wrappers* têm como objetivo extrair informações em textos estruturados ou semiestruturados, nos quais o processamento de linguagem é difícil de ser realizado, como em exemplo, uma Fatura de Telefonia Móvel Corporativa. Este trabalho fará uso desta técnica, baseando-se nas regras de formatação, delimitadores e tipografia das palavras presentes nos documentos.

2.2.4 Técnicas para Extração de Informação

Uma vez conhecidos os conceitos básicos sobre EI e as abordagens para a construção desse tipo de sistema, as principais técnicas utilizadas pelos *Wrappers* para extração da informação do texto serão apresentadas.

As técnicas definirão como o sistema implementará a função que preencherá o formulário final (tabela), e determinará que tipo de regra de extração deverá ser codificada manualmente pelo engenheiro do conhecimento ou aprendida de forma automática. Abaixo se encontram descritas as técnicas: Autômatos Finitos, Casamento de Padrões e Classificadores.

Autômatos Finitos

Um autômato finito ou máquina de estados finitos, é um modelo computacional que consiste em um conjunto de estados S, um estado inicial s e um conjunto de transições entre os estados que ocorrem com a leitura de um símbolo de um alfabeto Σ.

O autômato começa a realizar seu processamento a partir de seu estado inicial, e conforme os símbolos vão sendo lidos, mudanças de estados ocorrem dependendo da função de transição.

Formalmente, segundo [Hopcroft e Ullman 1979], define-se um autômato finito por uma 5-tupla (S, Σ, s_0, T, F), onde:

- S é um conjunto finito de estados;
- Σ é um alfabeto finito de símbolos de entrada;
- T é a função de transição ($T : S \text{ X } \Sigma \rightarrow S$);

- $s_0 \in S$, é o estado inicial;
- $F \subseteq S$ é o conjunto de estados finais.

Esta técnica é usada de diversas formas na computação, como a criação de analisadores léxicos, para a criação de uma linguagem de programação e seu compilador, editores de texto, para substituir padrões e também em modelagem de agentes em jogos eletrônicos.

Wrappers também podem utilizar esse tipo de técnica para realizar a extração de informação, a fim de determinar se uma dada cadeia de caracteres de um documento pode ou não preencher um determinado campo de um formulário de entrada, com base no processamento do autômato terminar ou não em um estado final.

De maneira geral, [Silva 2004] mostra que essa técnica pode ser usada para extrair informações de um documento, definindo:

- Os estados que aceitam os símbolos a serem extraídos, a fim de preencher um formulário de saída;
- Os estados que irão apenas consumir os símbolos irrelevantes encontrados no documento;
- Os símbolos do documento de entrada que provocaram a transição de um estado para outro.

Casamento de Padrões

Vários sistemas de EI baseiam-se em técnicas de casamento de padrões para realizar a extração de informação de textos. Os padrões de extração podem ser descritos por expressões regulares, que são mais intuitivas de escrever que os autômatos [Soderland et al. 1995].

O processo de extração se dá quando se realiza o casamento dos padrões com o texto de entrada, podendo ser utilizado tanto para textos estruturados, semiestruturados ou livres. Atualmente, existem diversas linguagens de programação que possuem expressões regulares, como Phyton, Java, JavaScript, que auxiliam o usuário na sua utilização e que podem ser usadas para a criação de sistemas de EI.

Classificadores

Um classificador é capaz de classificar uma entrada de texto, representado por um vetor de características, em uma categoria ou classe predefinida. Os problemas da extração de informação podem ser resolvidos por classificadores, de forma diferente. Dado um documento, não se deseja classificá-lo em uma determinada categoria, e sim extrair dele alguns trechos que preenchem um formulário de saída desejado. Sendo assim, o algoritmo recebe do documento de entrada um fragmento de texto e determina o grau de confiança que esse fragmento preenche corretamente cada um dos campos do formulário de saída.

Assim, para realizar a extração de informação com classificadores, é necessária uma heurística adequada para gerar fragmentos de texto candidatos a serem extraídos e um classificador, capaz de decidir qual campo cada fragmento deve preencher.

A geração de fragmentos relativamente significativos e corretos que serão utilizados é uma das maiores limitações para se tratar em problemas de EI com classificadores [Kushmerick e Thomas 2003], pois algumas vezes não é possível realizar a separação adequada do documento de entrada em cadeias de palavras candidatas a preencher algum campo do formulário de saída. A partir dos fragmentos, pode-se extrair uma série de características que formarão o vetor usado pelo classificador. As quais podem ser extraídas dos conhecimentos do especialista no domínio ou de processos de seleção automática [Yang e Pedersen 1997].

Classificadores buscam suprir a necessidade de se classificar instâncias de forma mais rápida e menos custosa, usando apenas as características dos objetos como critério para classificação, podendo possuir dois modelos, discriminativo ou probabilístico [Costa et al. 2007].

No modelo discriminativo, um padrão de classificação é usado para, a partir dos atributos, encontrar a classe do objeto; já no modelo probabilístico, é calculada uma probabilidade de o objeto pertencer a cada uma das classes [Hand, Mannila e Smyth 2001].

Os classificadores Bayesianos são exemplo de classificadores probabilísticos [Silla e Freitas 2011]. Nesse modelo caso os antecedentes sejam verdadeiros, os consequentes tem uma determinada probabilidade de acontecer.

O algoritmo Bayesiano é uma técnica estatística baseada no teorema de Thomas Bayes, e com ela é possível encontrar a probabilidade de certo evento ocorrer, dada a probabilidade de outro evento que já ocorreu, como mostra a Equação 2.1. O algoritmo parte do princípio de que

não existe relação de dependência entre os atributos.

$$Probabilidade(B|A) = \frac{Probabilidade(A|B) * Probabilidade(B)}{Probabilidade(A)} \quad (2.1)$$

Segundo [Zhang 2004], devido à sua simplicidade e alto poder preditivo os algoritmos Bayesianos, chamados de Naïve Bayes, comparados em seu trabalho, com os métodos de Árvore de Decisão e Redes Neurais Artificiais obtiveram resultados compatíveis. Tendo em vista sua fácil implementação e resultados satisfatórios, o algoritmo de Naïve Bayes será utilizado nesse trabalho.

Deve-se lembrar que o tipo de classificador criado depende da forma como as classes das instâncias se relacionam entre si, portanto, podem ser hierárquicos ou não hierárquicos.

2.2.5 Classificação Hierárquica

Os itens presentes em uma base de dados podem pertencer a classes que independem umas das outras ou podem ter uma relação hierárquica de dependências entre si. Essa existência de relação entre as classes definirá o tipo de classificador usado para rotular novos exemplos dos itens em questão. Para os casos em que não há relação entre as classes, o classificador usado é chamado de não hierárquico. Para os outros casos, o classificador é chamado hierárquico [Silla e Freitas 2011].

Em problemas hierárquicos, a organização das classes pode ser definida em dois tipos de estruturas: um grafo cíclico ou um grafo acíclico (árvore). Uma árvore pode ser representada por uma hierarquia de classes na qual cada classe possui somente uma classe pai. Já a estrutura de grafo cíclico, a classe filha pode ser associada a diversas classes pai, como nas Figuras 2.3 e 2.4.

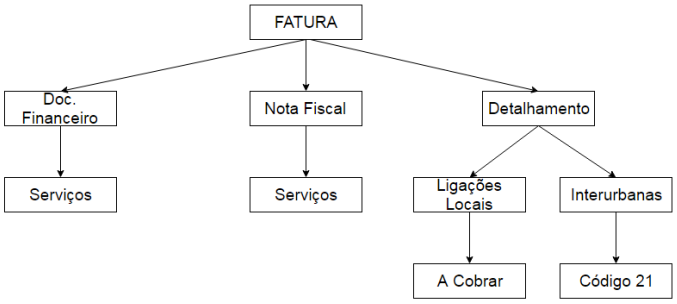


Figura 2.3: Estrutura hierárquica acíclica (grafo acíclico)

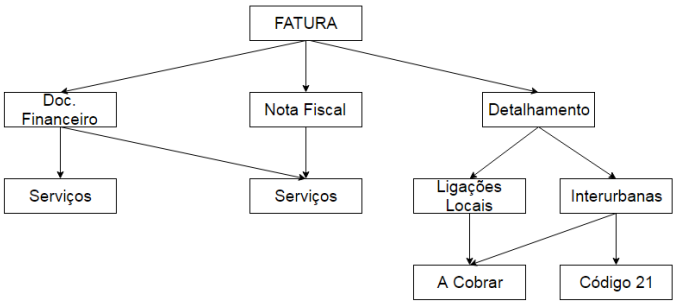


Figura 2.4: Estrutura hierárquica cíclica (grafo cíclico)

As abordagens para a resolução de problemas hierárquicos se diferenciam na forma em que a estrutura hierárquica é explorada, seja na simplificação da hierarquia, na utilização de um conjunto de classificadores planos tradicionais, ou na construção de um único classificador que considera toda a hierarquia de classes. Abaixo será apresentada uma breve introdução referente para cada uma dessas abordagens.

Abordagem de Classificação Plana

Essa abordagem de classificação também pode ser denominada na literatura como uma abordagem direta (*Direct Approach*) [Burred e Lerch 2003], e simplifica o problema de classificação hierárquica, de modo que o transforma em um problema de classificação tradicional, ignorando sua hierarquia de classes. Geralmente nessa abordagem, somente as classes folhas são consideradas no treinamento e no processo de classificação, como pode ser visto na Figura 2.5.

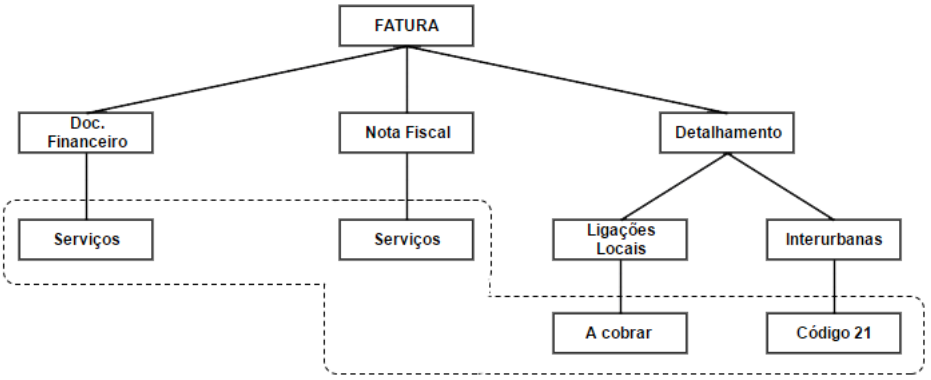


Figura 2.5: Abordagem de classificação plana

A vantagem dessa abordagem é a sua simplicidade de aplicação, pois um único classificador plano tradicional pode ser usado para prever um conjunto de classes hierárquicas. Entretanto, tem-se a desvantagem de não conseguir explorar as informações que são provenientes das relações de descendência na hierarquia [Paes 2012].

Abordagens por Classificadores Locais

As abordagens por classificadores locais visam explorar a hierarquia de classes através de uma perspectiva local, combinando classificadores que consideram isoladamente diferentes partes da hierarquia [Paes 2012]. Esses classificadores são categorizados pela forma em que as informações locais são exploradas: local por nó, local por nó pai e local por nível [Silla e Freitas 2011].

- **Abordagem de Classificação por Nó:** Essa abordagem local consiste no treinamento de um classificador associado a cada nó da hierarquia de classes, exceto o nó raiz, e após o treinamento, uma hierarquia de classificadores planos é formada. A Figura 2.6 mostra um exemplo desse tipo de abordagem, com a representação dos classificadores por meio de pontilhados em cada nó da árvore. Essa abordagem identifica por meio de cada classificador se a instância pertence ou não a classe à qual está associada.

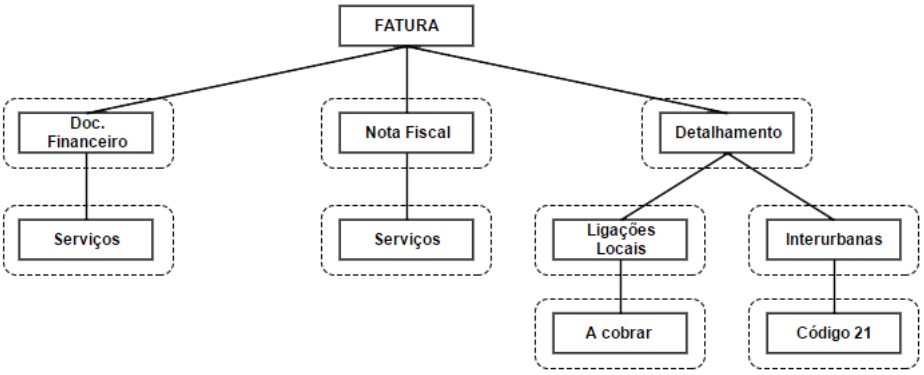


Figura 2.6: Abordagem de classificação local por nó

- **Abordagem de Classificação Local por Nó Pai:** Essa abordagem consiste em treinar um classificador plano tradicional para cada classe pai da hierarquia. Esses classificadores consideram somente suas classes filhas. A Figura 2.7 apresenta um exemplo desse tipo

de hierarquia, com a representação das classes pais através de pontilhados em cada nó da árvore, significando que essas possuem classificadores associados.

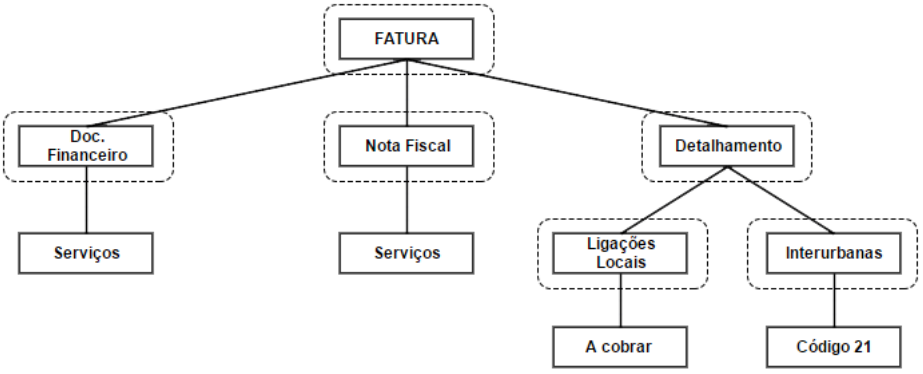


Figura 2.7: Abordagem de classificação por nó pai

Essa abordagem se baseia na hipótese de que, quando utilizados diferentes algoritmos de classificação em cada nó pai da hierarquia, a acurácia preditiva é melhorada[Secker et al. 2007], visto que cada nó da árvore pode apresentar um comportamento e informações diferentes, que quando utilizado individualmente, não gera conflito entre eles.

- **Abordagem de Classificação Local por Nível:** Consiste no treinamento de um classificador plano para cada nível da hierarquia de classes. Em cada camada da hierarquia, um classificador fica responsável por selecionar uma das classes existentes em cada nível, como na Figura 2.8.

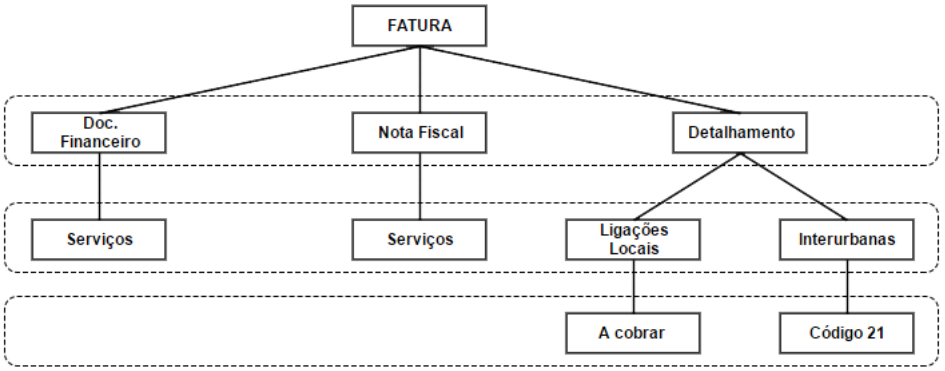


Figura 2.8: Abordagem de classificação local por nível

Para que essa abordagem seja utilizada, faz-se necessário um tratamento posterior para corrigir possíveis classificações inconsistentes nas diferentes camadas da hierarquia, pois o principal problema aqui é a ocorrência de inconsistências nos resultados obtidos pelos classificadores.

- **Abordagem de Classificação Global:** Essa abordagem, que também pode ser encontrada na literatura com a denominação de abordagem *big-bang* [Carvalho 2011], consiste na construção de um único modelo que considera toda a hierarquia de classes, como pode ser visto na Figura 2.9, que apresenta um único classificador representado por um pontilhado entorno de toda a árvore.

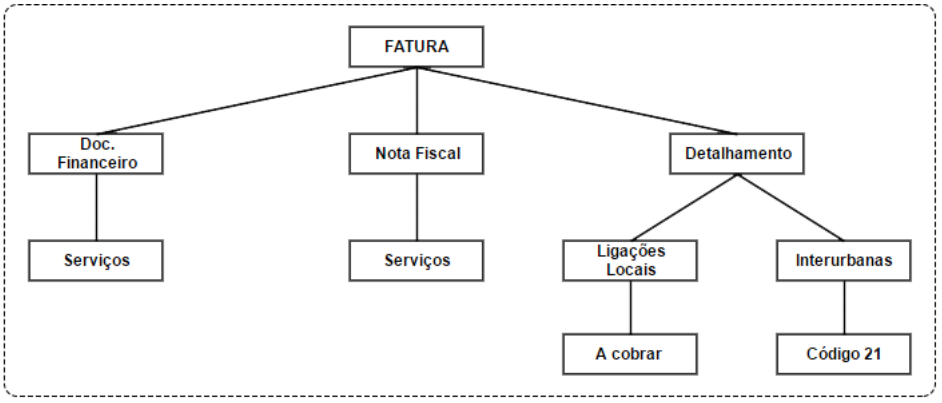


Figura 2.9: Abordagem de classificador global

A principal característica dessa abordagem é a sua capacidade de considerar todas as

classes da hierarquia em um único treinamento. Como consequência, a abordagem global sofre com a falta de modularização, gerando um modelo com maior complexidade [Paes 2012].

2.2.6 Características Hierárquicas em Textos

Estruturas hierárquicas de dados presentes em textos são encontradas com frequência em diversos estudos. Elas são caracterizadas pela presença de unidades agrupadas em outras unidades maiores, que por sua vez podem ou não formar novos grupos.

Para [Jubran, Urbano e Risso 1992], em um plano ou estrutura hierárquica, as sequências textuais se desdobram em supertópicos e subtópicos, dando origem a quadros tópicos, caracterizados, obrigatoriamente, pela concentração em um tópico mais abrangente e pela divisão interna em tópicos co-constituintes e, possivelmente, por subdivisões sucessivas no interior de cada tópico, de forma que esses podem vir a ser ao mesmo tempo supertópicos ou subtópicos, mediante uma relação de dependência entre dois níveis não imediatos.

Este tipo de estrutura é bastante comum em faturas telefônicas, que apresentam em seu conteúdo detalhamento de números que estão dentro de grupos maiores, sendo possível determinar o tipo de detalhamento apenas se esses grupos forem reconhecidos. Este tipo de estrutura pode servir de auxílio para os sistemas de EI, visto que conhecida previamente sua organização é possível utilizar uma abordagem como a Classificação por Nó, para determinar a classe de cada um dos dados presentes nos documentos.

2.3 Trabalhos Correlatos

Esta seção trata de trabalhos relacionados à classificação hierárquica e extração de informações em documentos textuais, uma vez que o objetivo deste trabalho é o desenvolvimento de uma aplicação capaz de extrair informações relevantes de faturas de telefonia móvel corporativa, que se encontram em formato textual e possuem em sua organização interna dependências de tópicos.

Diversos métodos e abordagens para a resolução destes problemas de classificação hierárquica podem ser encontrados na literatura, como é o caso do trabalho de [Paes 2012], que apresenta novas estratégias para a classificação hierárquica local por nível.

Exemplos de problemas que possuem classes organizadas hierarquicamente podem ser encontrados em diversas áreas de aplicação, como é o caso da Bioinformática, onde existem importantes trabalhos que visam classificar proteínas em classes funcionais, que se encontram organizadas hierarquicamente, como ocorre nos trabalhos de [Vens et al. 2008] e [Menezes 2014]. Na área de classificação de documentos, textos podem ser caracterizados por sua hierarquia de assuntos ou tópicos como em [Sun e Lim 2001] e [Silla e Freitas 2011]. Em aplicações de reconhecimento de imagens, objetos também podem ser classificados por suas relações de dependência, como nos trabalhos de [Rossoni, Pires e Zanotta 2013] e [Vens et al. 2008].

As pesquisas na área de EI, tinham como principal motivação povoar automaticamente base de dados [Jacobs e Rau 1993], como é o caso deste trabalho. Porém, sistemas de EI também permitem melhorar o desempenho de sistemas que utilizam a recuperação de informação, evitando, por exemplo, a ocorrência de redundâncias em textos que tratam de assuntos semelhantes. Pelo estudo bibliográfico realizado pode-se notar a grande utilização da EI no contexto de obtenção de informações referentes a artigos científicos, como é o caso dos trabalhos de [Álvarez 2007] e de [Silva 2004], que induzem de forma automática, um conjunto de regras para a extração de informações de artigos científicos. Essas informações estão presentes no corpo dos artigos e em suas referências bibliográficas, e visam como objetivo transformar os documentos em um formato estruturado, mapeando as informações contidas nos artigos para uma estrutura tabular.

Na literatura, podem ser encontrados vários trabalhos relacionados à extração de informações de resumos, referências bibliográficas ou de artigos científicos, no entanto, não foi possível encontrar trabalhos referentes a EI relacionadas com a área de telecomunicações, mais precisamente, em faturas de telefonia móvel corporativa.

2.4 Considerações Finais

Como apresentado neste capítulo, existem várias técnicas que podem ser utilizadas na extração de informação, cada uma apresentando vantagens e desvantagens, mostrando-se mais ou menos adequadas a cada tipo de problema. A Tabela 2.1 indica quais técnicas são indicadas para cada tipo de texto.

Tabela 2.1: Técnicas de EI para cada tipo de texto

Técnica	**Texto estruturado**	**Semiestruturado**	**Não estruturado**
Casamento de Padrões	X	X	X
Autômatos	X	X	
Classificadores		X	
PLN		X	X

As técnicas que envolvem casamento de padrões são mais fáceis de serem escritas manualmente em comparação com os autômatos finitos. Além disso, são adequadas para extrair informações dos três tipos de textos apresentados, desde que exista uma regularidade no texto a ser extraído, e que essa possa ser representada por meio de uma expressão regular. Em contrapartida, os autômatos são mais adequadamente construídos por aprendizagem de máquina, além de serem capazes de extrair informações de textos estruturados e alguns textos semiestruturados.

As técnicas baseadas em classificadores são capazes de extrair informação de textos semiestruturados, tendo em vista que o texto não deve possuir um número muito grande de fragmentos candidatos a preencher alguns dos campos do formulário de saída. Como essas técnicas podem fazer uso de um grande número de características dos fragmentos do texto para realizar a classificação, apresentam a desvantagem de obter resultados ótimos apenas localmente, por realizarem a classificação independente para cada fragmento.

Capítulo 3

Modelo Proposto

Este trabalho tem por objetivo propor um modelo para a resolução do problema de Extração de Informação em documentos cujo conteúdo relaciona-se de forma hierárquica, fundamentado no processo de Mineração de Texto discutido no Capítulo 2, como apresentado na Figura 3.1.

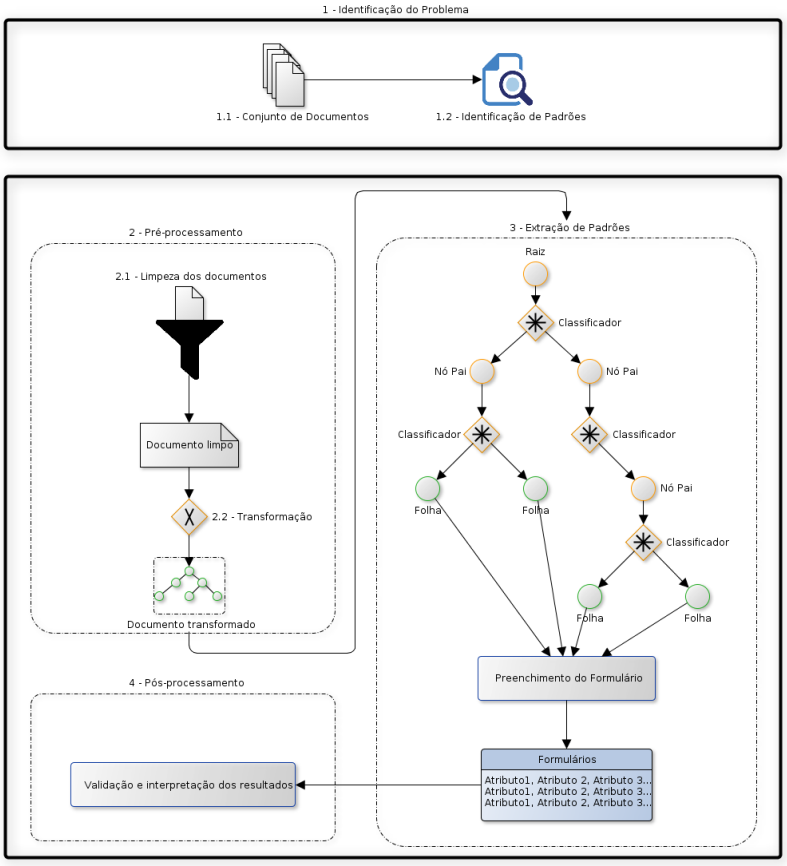


Figura 3.1: Modelo adaptado para elaboração do trabalho
Fonte: O autor

Esse modelo é divido em dois processos, sendo eles: a Identificação do problema e a Extração de Informações. O primeiro processo tem como objetivo principal a coleta de amostras de documentos a serem utilizados (Etapa 1.1), bem como o estudo das características presentes no conteúdo da amostra e a definição de um formulário de saída capaz de armazenar as informações pertinentes extraídas dos documentos (Etapa 1.2).

O estudo das características do texto dentro de um documento possibilita a identificação das técnicas a serem utilizadas ao longo do processo proposto e o projeto do formulário final a ser preenchido na etapa de Extração de Padrões. Após ter sido realizada a Identificação do problema e suas etapas, o modelo proposto segue com o processo Extração de Informações. Esse processo engloba as etapas de Pré-processamento (Etapa 2), na qual se realiza a limpeza (Etapa 2.1) e transformação (Etapa 2.2) do conteúdo do documento, a fim de impor uma condição inicial para a aplicação da técnica escolhida na etapa de Extração de Padrões (Etapa 3).

Na etapa de Extração de Padrões aplica-se a Classificação Hierárquica como técnica para EI, a fim de preencher o formulário elaborado no primeira processo deste modelo. Ao final, na etapa de Pós-processamento (Etapa 4), os resultados são avaliados com o objetivo de validar a técnica utilizada.

3.1 Identificação do Problema

A coleta de documentos foi realizada mediante a disponibilização de uma amostra de documentos por parte de um especialista no domínio do problema. Após a coleta, um estudo foi realizado sobre os documentos coletados. A principal característica encontrada nestes documentos refere-se à relação hierárquica existente entre os textos, ou seja, a extração de informações de um determinado trecho do documento depende do contexto onde ele se encontra. É também na etapa de Identificação do Problema onde o formulário a ser preenchido pela etapa de Extração de Padrões é projetado, uma vez que o principal objetivo neste momento é entender os padrões do conteúdo do documento a fim de encontrar uma solução para estruturá-lo.

Este estudo é fundamental para a identificação das técnicas mais adequadas às etapas do processo seguinte.

3.2 Pré-processamento

As técnicas e padrões investigados no processo anterior resultaram na elaboração de um pré-processamento dividido em duas subetapas: (i) limpeza dos documentos e (ii) transformação.

A limpeza dos documentos consiste na eliminação de conteúdos que sejam irrelevantes para a tarefa de Extração de Informação, de modo a considerar apenas os trechos úteis para a estruturação do formulário final.

Para realizar a limpeza, é necessário que antes o documento seja convertido para um formato ao qual seja possível a manipulação do texto sem qualquer estilo. Desta forma, torna-se menos custosa a aplicação de técnicas computacionais. Esta conversão é necessária apenas em alguns tipos de arquivos, a exemplo daqueles em formato PDF. Para tal conversão estão disponíveis várias ferramentas no mercado. Pelo uso popular, optou-se neste trabalho pela utilização do iText [Lowagie 2010].

```
Detalhamento de ligações e serviços do celular([illegible])
AEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEB
Mensalidades e Pacotes Promocionais
CEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEED
Total (R$) Descrição
Assinatura Plano Sob Medida 5,61
   Desconto Assinatura Plano Sob Medida -0,42
Consumo Compartilhado 0,00
AEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE AE
Total R$ 5,19
F F
CEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE CE
Pág.       5 /      18

-------------------------------------


--Página Nro -----6-------

Detalhamento de ligações e serviços do celular([illegible])
AEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEB
Mensalidades e Pacotes Promocionais
CEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEED
Total (R$) Descrição
Assinatura Plano Sob Medida 5,61
   Desconto Assinatura Plano Sob Medida -0,42
Consumo Compartilhado 0,00
Gestor Online - Controle Completo 5,08
Pacote Ilimitado Internet 50MB 12,37
   Desconto Pacote Ilimitado Internet 50MB -0,92
Serviço Tarifa Zero 4,65
   Desconto Serviço Tarifa Zero -0,35
AEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE AE
Total R$ 26,02
F F
CEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE CE
```

Figura 3.2: Trecho de uma fatura antes da limpeza
Fonte: O autor

A título de ilustração, a Figura 3.2 apresenta, já no formato de texto plano, um trecho de

uma fatura telefônica antes da limpeza, onde existem vários trechos irrelevantes para o processo de Extração de Informação, a exemplo do trecho ("Pag. 5 / 18"), que representa o rodapé com a informação da página atual do documento. Pretende-se com a limpeza, eliminar este tipo de conteúdo.

Para ilustrar o comportamento desta operação, a Figura 3.3 apresenta o resultado da limpeza para o exemplo da Figura 3.2.

```
Detalhamento de ligações e serviços do celular([illegible])
Mensalidades e Pacotes Promocionais
Assinatura Plano Sob Medida 5,61
   Desconto Assinatura Plano Sob Medida -0,42
Consumo Compartilhado 0,00
Total R$ 5,19
Detalhamento de ligacões e servicos do celular([illegible])
Mensalidades e Pacotes Promocionais
Assinatura Plano Sob Medida 5,61
   Desconto Assinatura Plano Sob Medida -0,42
Consumo Compartilhado 0,00
Gestor Online - Controle Completo 5,08
Pacote Ilimitado Internet 50MB 12,37
   Desconto Pacote Ilimitado Internet 50MB -0,92
Serviço Tarifa Zero 4,65
   Desconto Serviço Tarifa Zero -0,35
Total R$ 26,02
```

Figura 3.3: Trecho de uma fatura depois da limpeza
Fonte: O autor

Esta operação, como discutida por [Hand, Mannila e Smyth 2001], pode ser realizada com o auxílio de um especialista no domínio, tanto para a identificação de características utilizadas em um processo de aprendizado de máquina quanto para a identificação de regras utilizadas em um processo manual. Independente da forma como a operação será realizada, o resultado deve representar exatamente aquilo que se pretende estruturar.

Com o texto limpo, uma transformação é necessária para impor uma estrutura inicial, essencial para a aplicação da tarefa de Extração de Informação, que será realizada na etapa seguinte.

Esta estrutura é representada por meio de uma árvore e pode ser alcançada explorando-se a relação hierárquica existente no texto.

No exemplo da Figura 3.3, a relação hierárquica pode ser observada da seguinte forma: Os títulos são considerados nós intermediários da árvore e os nós folhas são identificados pelos trechos que contém alguma informação monetária. Seguindo esta lógica, o resultado da transformação para esse exemplo é mostrado na Figura 3.4. Após a identificação/delimitação do problema e transformação dos textos, o modelo avança para a etapa de Extração de Padrões.

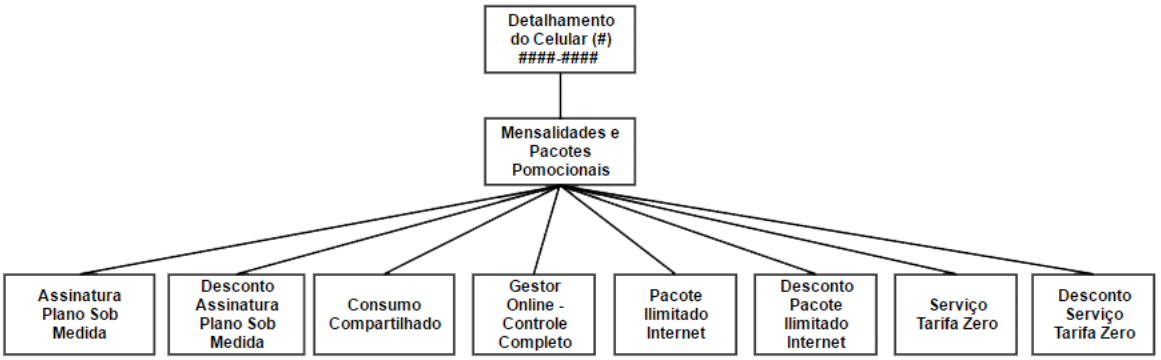


Figura 3.4: Resultado da transformação do trecho da fatura
Fonte: O autor

3.3 Extração de Padrões

Sabendo que o documento já está transformado, e conhecendo que sua estrutura interna é composta por uma hierarquia de tópicos, uma das técnicas de classificação hierárquica apresentadas na Seção 2.2.5, pode ser utilizada para que se consiga manter e extrair corretamente as informações sem comprometer a estrutura interna dos documentos, a exemplo da abordagem de classificação hierárquica por nó pai.

Nesta abordagem, que consiste no treinamento de um classificador associado a cada nó da hierarquia de classes, pode-se utilizar o classificador bayesiano em cada um dos nós da hierarquia. Com o objetivo de classificar um trecho da fatura, a classificação realiza uma busca em profundidade na árvore de classificadores.

Suponha que se deseja classificar o item da fatura "Assinatura plano sob medida 5,61", a exemplo da Figura 3.5, os passos realizados pela abordagem de classificação hierárquica por nó pai para a classificação deste item serão apresentados.

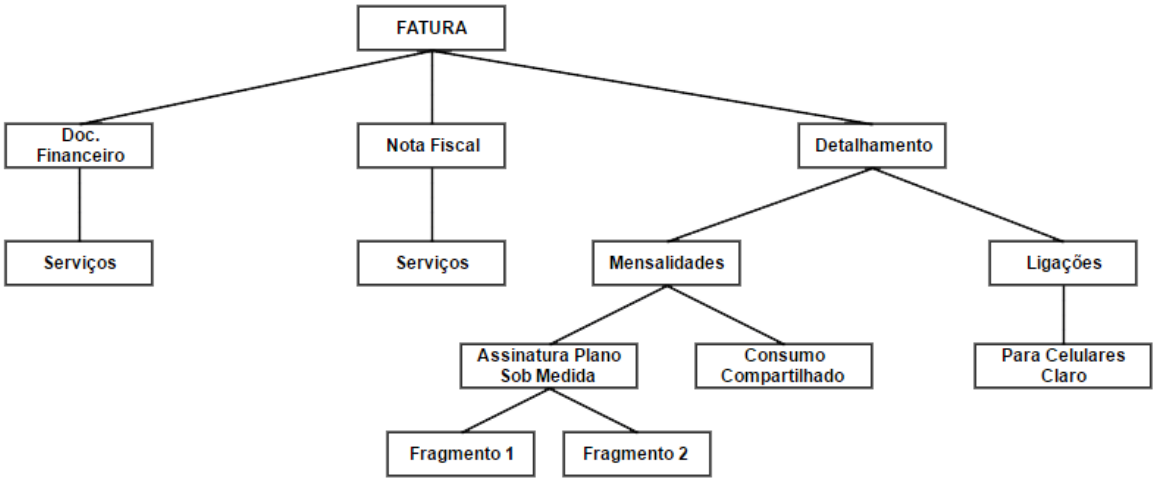


Figura 3.5: Classificação do item da fatura
Fonte: O autor

O trecho a ser analisado estará presente na raiz da árvore, que contém um classificador responsável por identificar um Documento Financeiro, uma Nota Fiscal ou um Detalhamento de Número. Tendo em vista que o classificador bayesiano possui conhecimento de um item que já foi previamente classificado, o item analisado será classificado por este nó como Detalhamento.

Sabendo que a classe é Detalhamento, o nó raiz enviará o item para este classificador, o qual será responsável por classificar o item entre Mensalidades e Ligações. O item então será classificado como Mensalidades, devido ao fato do classificador deste nó também possuir conhecimento de outros itens já classificados. O mesmo acontece com o nó Mensalidades, que decidirá entre as classes Plano Sob Medida ou Consumo Compartilhado.

Finalmente o item analisado é classificado como Plano Sob Medida, e tem seu texto fragmentado entre dois nós folhas, identificados na Figura acima como (Fragmento 1) e (Fragmento 2). O Fragmento 1 obtém o trecho "Assinatura Plano Sob Medida"e o Fragmento 2 obtém o trecho "5,61".

Ao final da classificação hierárquica, o formulário de saída poderá ser preenchido, como mostrado na Tabela 3.1. Tendo o formulário de saída sido preenchido, a etapa de Pós-processamento analisará se o resultados condizem exatamente com o que se apresenta na fatura.

Tabela 3.1: Exemplo de formulário de saída preenchido

Número da Linha	Tipo	Modalidade	Descrição	Valor
X	Assinatura	VOZ	Plano Sob Medida	5,61

3.4 Pós-processamento

Essa etapa é responsável pela validação e avaliação do conhecimento extraído. A avaliação é realizada de forma subjetiva, utilizando o conhecimento de um especialista no domínio, que verificará se a estrutura resultante no processo condiz com as informações presentes nos documentos, e de forma objetiva por meio de validações *ad hoc*, por meio de índices estatísticos que indicam a qualidade dos resultados extraídos.

3.5 Considerações Finais

Este capítulo apresentou o modelo elaborado para a resolução do problema de Extração de padrões, apresentado os processos necessários para o seu desenvolvimento. Nele também foram apresentadas as técnicas e padrões investigados para realizar as subetapas de limpeza e transformação dos documentos.

Nota-se que a limpeza é um processo relativamente importante, considerando que os documentos apresentam uma grande quantidade de informações que não são relevantes à extração. Isso faz com que essas instâncias não se propaguem para as próximas etapas, minimizando o custo operacional do sistema. Já a transformação é essencial para impor uma estrutura inicial, visto que a técnica aplicada na extração de informação será a classificação hierárquica.

Capítulo 4

Construção do Modelo

A importância da gestão dos serviços de telecomunicações dentro de uma empresa vão do acompanhamento do uso dos serviços à redução de custos para a empresa. Dada a complexidade na mecânica e comercialização destes serviços, realizar este trabalho manualmente pode ser ainda mais custoso, sobretudo para grandes empresas [Cavalcante 2009].

Pretendendo auxiliar na resolução de problemas como os descritos por [Cavalcante 2009], tomando como base o modelo proposto no capítulo anterior e, considerando que uma fatura telefônica contém inúmeras páginas de detalhamentos, fica praticamente inviável realizar a verificação de informações manualmente.

Este capítulo tem como objetivo apresentar os passos realizados para a construção de um sistema computacional capaz de extrair informações de faturas de telefonia móvel corporativa por meio do uso de classificadores hierárquicos. Será apresentado o processo que é baseado na obtenção do *corpus* de faturas, a explanação das características destes documentos, definição do *corpus* de faturas para a realização deste processo, bem como o formulário de saída esperado.

4.1 Identificação do Problema

O estudo das características presentes nas faturas de telefonia móvel corporativa foi realizado sobre um *corpus* de 1200 faturas da operadora Claro Telecom Participações S.A., disponibilizadas pela empresa Orb it Sistemas, situada na cidade de Cascavel-PR. A coleta destes documentos foi realizada mediante um termo de sigilo e confidencialidade, em razão da proteção de informações confidenciais dos clientes presentes nas faturas, de modo que nenhuma informação relacionada a dados pessoais de clientes e fornecedores, a exemplo de, mas não

restritos a, nome e endereço, serão utilizadas, tampouco divulgadas. Por esse motivo algumas imagens deste trabalho tem seu conteúdo borrado.

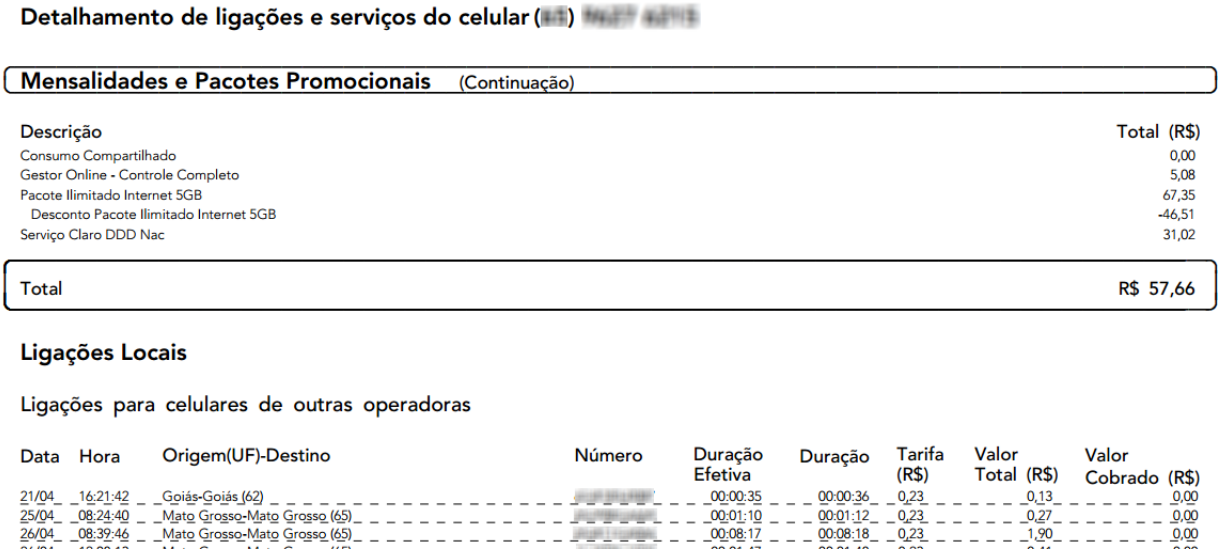
Optou-se por trabalhar com apenas uma operadora, devido ao fato de que o padrão de organização dos dados de uma fatura podem diferenciar de uma operadora para outra. A construção do modelo proposto no Capítulo 3 fundamentou-se no estudo das características contidas nas faturas obtidas, discutidas ao longo da próxima seção.

Características das Faturas

Para [Cavalcante 2009], a fatura telefônica trata-se de um documento complexo pelas seguintes razões:

1. São excessivamente técnicas;
2. Tem um padrão próprio;
3. Estão enquadradas em uma legislação específica;
4. Refletem a complexa estrutura da operação.

Estas razões justificam a escolha de apenas uma operadora para a construção do modelo proposto, haja vista cada operadora ter o seu próprio formato. O formato dessas faturas apresenta, na prática, uma grande variação na sua estrutura, o que torna a extração de informação neste domínio uma tarefa bastante trabalhosa. A partir de uma fatura telefônica móvel corporativa, podem ser extraídas diversas informações, como valores, horários, números de origem e destino, tipo de ligação, como mostra a Figura 4.1.



Detalhamento de ligações e serviços do celular ([illegible]**)**

Mensalidades e Pacotes Promocionais (Continuação)

Descrição	Total (R$)
Consumo Compartilhado	0,00
Gestor Online - Controle Completo	5,08
Pacote Ilimitado Internet 5GB	67,35
Desconto Pacote Ilimitado Internet 5GB	-46,51
Serviço Claro DDD Nac	31,02
Total	R$ 57,66

Ligações Locais

Ligações para celulares de outras operadoras

Data	Hora	Origem(UF)-Destino	Número	Duração Efetiva	Duração	Tarifa (R$)	Valor Total (R$)	Valor Cobrado (R$)
21/04	16:21:42	Goiás-Goiás (62)	[illegible]	00:00:35	00:00:36	0,23	0,13	0,00
25/04	08:24:40	Mato Grosso-Mato Grosso (65)	[illegible]	00:01:10	00:01:12	0,23	0,27	0,00
26/04	08:39:46	Mato Grosso-Mato Grosso (65)	[illegible]	00:08:17	00:08:18	0,23	1,90	0,00

Figura 4.1: Informações presentes em faturas telefônicas

O que há de comum entre as faturas telefônicas é o seu objetivo de apresentar e detalhar os gastos dos serviços de telecomunicações utilizados dentro de um determinado período. Estes gastos são categorizados em dois grupos, gastos fixos e gastos variáveis.

Os gastos fixos estão associados aos serviços contratados. Neste contexto da fatura estão descritos os valores e franquias, bem como o período de vigência destes serviços. Já os gastos variáveis estão associados a consumos não previstos na contratação e que são tarifados de forma avulsa, gerando um gasto acima do contratado.

A fim de obter mais detalhes sobre estes gastos, a maior parte da fatura destina-se ao detalhamento de cada item consumido. Um item pode ser uma ligação, uma mensagem ou um tráfego de internet e o detalhamento de cada item traz informações como horário em que o item foi consumido, número que o consumiu, origem e destino do consumo, valor cobrado, dentre outros.

A extração das informações contidas nestes detalhamentos possui como principal motivação a criação de uma base de dados capaz de fornecer uma ferramenta útil para serviços de auditoria e consultoria na área de Gestão de Custos em Telecomunicações, trazendo mais praticidade e qualidade, uma vez que o profissional da área não precisaria consultar e analisar tais informações de forma manual.

A fim de escolher a técnica mais adequada à tarefa de Extração de Informação, é necessário

antes investigar as características do texto presentes em uma fatura telefônica.

Tipo e características do texto

O texto de uma fatura telefônica é do tipo semi-estruturado e possui as seguintes características:

- Campos ausentes: As seções das faturas telefônicas, em geral, não possuem informações para preencher todos os campos do formulário de saída. Por exemplo, o campo tarifa encontrado em descrições de utilização pode não aparecer em determinados itens presentes na fatura.

- Estilo telegráfico: Muitas das palavras que aparecem em um título de ligação estão abreviadas. Por exemplo: Recebidas (Rec.), Longa Distância (LD.) e etc.

- Ausência de delimitadores precisos: Não existem delimitadores que definam com certeza onde começa e onde termina cada campo a ser extraído da fatura. Alguns delimitadores, como o espaço entre os itens, também podem aparecer no meio de um campo a ser extraído, como ocorre muitas vezes com o campo "Origem(UF)-Destino"que pode conter nomes de cidades compostos, como "São Paulo".

- Hierarquia de tópicos: As faturas telefônicas apresentam em seu conteúdo uma hierarquia de itens que devem ser levados em consideração para que se consiga obter corretamente a informação desejada. Títulos de ligações locais por exemplo podem contem subtítulos como Ligações a cobrar, que devem ser e extraídas em conjunto, para que se consiga obter a informação corretamente.

Todas essas características foram determinantes para a elaboração do modelo proposto. Em especial a característica de Hierarquia de tópicos, que motivou a utilização da técnica de Classificação Hierárquica.

Conhecendo a finalidade de uma fatura telefônica, obtendo um domínio razoável a respeito do seu conteúdo e identificando as principais características do texto presente em uma fatura, é possível agora definir um formulário de saída que deverá ser preenchido no processo de Extração de Informação.

Definição do Formulário de Saída

O formulário de saída define as informações que podem ser extraídas a partir de uma fatura telefônica. Para definição deste formulário, utilizaremos os campos que possuem maior importância dentro das faturas, identificados e explicados na Tabela 4.1.

Tabela 4.1: Campos do formulário de saída a serem preenchidos pelo sistema de EI

Campo	**Descrição**
Tipo	Tipo do detalhamento apresentado (Plano; Pacote; Consumo; Geral; Desconto)
Descrição	Descrição do item extraída da própria linha (Nome de Plano; Nome de Serviço; Nome Descontos)
Modalidade	Modalidade utilizada(Voz; Dados; SMS; MMS; Outros Serviços; Geral)
Subtipo do Serviço	Especificação mais detalhada do Tipo(Local; Longa Distância; Internacional; Exterior; Serviço)
Data	Data do consumo
Hora	Hora do consumo
Tipo do Terminal	Tipo de aparelho utilizado (Fixo; Móvel; Rádio; Nenhum)
Deslocamento	Se a utilização foi feita em deslocamento,ou seja,fora de sua área de registro (Sim; Não)
A Cobrar	Se a utilização foi feita a cobrar (Sim; Não)
Grupo	Se a utilização foi feita para dentro do grupo, ou seja,para números do mesmo CNPJ, ou para os demais(Intragrupo; Extragrupo)
Realizada	Se a ligação foi realizada ou recebida (Realizada; Recebida)
CSP	O Código de Seleção de Prestadora utilizado (TIM-41; GVT-25)
Número do Acesso	Número ao qual a descrição se refere
Número Destino	Número que recebeu a ligação
Operadora	Se a ligação foi para um número de mesma operadora (Mesma; Outra)
Cobrança	O tipo de cobrança que foi realizada (Isento; Franquia; Exedente)
Utilização	Valor de utilização referente a unidade do consumo
Unidade	Unidade do consumo (Minutos; Envios; MegaBytes)
Tarifa	Valor referente a tarifa praticada do consumo
Valor Total	Valor total do serviço
Valor Cobrado	Valor cobrado pelo serviço

A técnica escolhida para realizar o pré-processamento usa a classificação plana (comum) e a técnica escolhida para realizar a tarefa de Extração de Informação usa a Classificação Hierárquica descrita no Capítulo 2.

Como já discutido, a classificação, tanto plana quanto hierárquica, são processos supervisionados e necessitam de um conjunto (*corpus*) etiquetado de faturas para treinar seus classificadores, a fim de alcançar o objetivo definido na tarefa a ser executada.

Antes de abordar a construção da etapa de pré-processamento e da tarefa de Extração de Informação, a Seção 4.2 apresenta o processo de etiquetagem do *corpus* de faturas selecionado para este trabalho.

4.2 Etiquetagem do Corpus de Faturas

Conforme dito na Seção 2.2.4, um classificador é capaz de classificar uma entrada de texto ou instância em uma etiqueta ou classe predefinida. Para tanto, é necessário treiná-lo com um conjunto de instâncias já etiquetadas. Neste trabalho, o processo de classificação da etapa de pré-processamento e da tarefa de extração de informação requer que a fatura telefônica seja dividida em várias instâncias de texto (linhas da fatura), que serão submetidas primeiramente à etapa de pré-processamento, para avaliar se as mesmas são relevantes ou irrelevantes, e posteriormente ao processo de extração de informação, a fim de extrair da instância as informações que deverão preencher o formulário de saída descrito anteriormente.

O *corpus* etiquetado de fatura utilizado neste trabalho, também compreendido como uma base de conhecimento, armazenará para cada instância, duas classes, denominadas Classe Estrutural e Classe de Informação, onde a primeira tem como objetivo orientar a etapa de pré-processamento e a segunda é usada para orientar a tarefa de extração de Informação.

A etiquetagem foi realizada utilizando engenharia do conhecimento e tomou como base o estudo realizado na identificação do problema. As ferramentas e tecnologias utilizas para a construção do sistema, bem como para realizar a etiquetagem do *corpus*, foram o SGBD PostgreSQL, utilizado para armazenar as informações no banco de dados, a linguagem de programação Java, bem como a biblioteca Itext, utilizada para construção, leitura e manipulação de documentos em formato PDF. As próximas seções apresentam o processo de definição das classes estruturais e de informação.

4.2.1 Definição das Classes Estruturais

As classes estruturais têm como objetivo representar a relevância e o nível hierárquico de cada uma das instâncias de uma fatura. A relevância é importante para a limpeza do documento, de modo a filtrar as instâncias que não são relevantes para a tarefa de Extração de Informação. Já o nível hierárquico é essencial para a etapa de transformação onde uma estrutura inicial em formato de árvore é imposta sobre as instâncias relevantes.

Neste contexto, as classes estruturais que proporcionam a realização do pré-processamento são:

- NIVEL-1: Instâncias que possuem um grau mais externo dentro da fatura, como é o caso

da linha "Detalhamento do Celular";

- NIVEL-2: Instâncias que identificam tópicos internos aos de NIVEL-1. Exemplo: "Serviços (Torpedos, Hits, Jogos, etc)";
- NIVEL-3: Instâncias que identificam tópicos internos aos de NIVEL-2. Exemplo: "Dados (MB)";
- FOLHA: Identifica as instâncias mais internas presentes nas faturas, ou seja, aquelas que contém descrições de contratação e de utilização.
- LIXO: Instâncias que apresentam características que as diferenciam das linhas importantes da fatura.

A distinção entre os itens de nível 1, 2 e 3 foram realizadas sem maiores complexidades, em razão da discrepância de características existentes entre eles e as outras instâncias presentes na fatura. Em contrapartida, houve dificuldade para diferenciar classes de nível folha e lixo, devido ao fato de se apresentarem de forma semelhante. Um exemplo dessa complexidade pode ser visualizada na Figura 4.2, onde é possível perceber que as linhas possuem similaridade, e para diferenciá-las é necessário o uso de características para sua identificação. Algumas das características utilizadas para discernir as linhas, são:

- As linhas que possuem o caractere $ foram consideradas não-relevantes para a extração de informações, visto que as linha relevantes não apresentam esse tipo de caractere.
- Linhas que apresentam as características de números de página e a palavra "(continuação)"foram consideradas irrelevantes para a extração, pelo fato de não apresentarem informações pertinentes.
- Itens como quebra de página, e símbolos que identificam objetos como retângulos ("FFF"), sites, códigos de barras, etc. foram identificados como "LIXO".
- Instâncias que apresentam dois espaços consecutivos, e instâncias que apresentam espaços no final de sua cadeia de caracteres, como visto respectivamente nas linhas 1 e 3 da Figura 4.2, também foram identificadas como "LIXO".

	Linha com Lixos	Linha Relevante
1	Ligações para celulares Claro #,# #,# #,# #,#	Ligações para celulares Claro #,# #,# #,# - #,#
2	Desconto Pacote Ilimitado Internet #GB - - - - R$ -#,# -	Desconto Pacote Ilimitado Internet #GB -#,# -#,# - -#,# - -#,#
3	Pacote Ilimitado Internet #GB #,#	Pacote Ilimitado Internet #GB #,#
4	Assinatura Plano Sob Medida R$ #,# R$ #,# R$ #,# R$ #,#	Assinatura Plano Sob Medida #,# #,# #,# - #,#

Figura 4.2: Amostra de similaridade entre linhas de lixo e folha

Após a compreensão das instâncias (linhas), as que obtiveram identificação de LIXO, foram descartadas das próximas etapas de EI.

4.2.2 Definição da Classe de Informação

Nesta etapa foi acrescentada uma etiqueta que corresponde a uma classe escolhida com base no tipo da linha. Essa etiqueta serve para identificar o tipo de informação que será preenchida no formulário de saída. Deve-se lembrar que essa etiqueta não identifica diretamente o campo a ser preenchido, e sim informações que podem ser usadas como base para preencher diversos campos. A explanação das etiquetas usadas para identificar as linhas são apresentadas na Tabela 4.2.

Tabela 4.2: Etiquetas usadas na etiquetagem das classes de informação

Etiqueta	Descrição
A COBRAR-RECEBIDA	Identifica que uma ligação foi recebida de forma a cobrar
CONSUMO	Consumos de forma geral que foram realizados na fatura
CONSUMO-EXTERIOR	Consumos realizados no exterior
CONSUMO-VOZ	Consumos do tipo voz
CONSUMO-VOZ-INTERNACIONAL	Consumos internacionais do tipo voz
CONSUMO-VOZ-LD	Consumos de longa distância do tipo voz
CONSUMO-VOZ_LOCAL	Consumos locais do tipo voz
CONSUMO-VOZ-MESMA OPERADORA	Consumos para a mesma operadora do tipo voz
CONSUMO-VOZ-SERVIÇO	Consumos de serviços do tipo voz
DADOS	Dados utilizados (MegaBytes)
DESLOCAMENTO	Ligações em deslocamento
DESLOCAMENTO-RECEBIDA	Ligações recebidas em deslocamento
DETALHAMENTO	Detalhamento de números de um celular
DOC-FINANCEIRO	Identificação de documentos financeiros
EXTRAGRUPO-MESMA OPERADORA	Ligações extragrupo para a mesma operadora
FIXO	Ligações feitas/recebidas de telefones fixos
INTRAGRUPO-MESMA OPERADORA	Ligações feitas para celulares do mesmo grupo e da mesma operadora
LD	Ligações longa distância
MENSALIDADES E PACOTES	Identificação de mensalidades e pacotes
MESMA OPERADORA	Itens utilizando a mesma operadora
MESMO CSP	Itens utilizando o mesmo código de seleção de prestadora
NOTA FISCAL	Identificação de notas fiscais
OUTRA OPERADORA	Itens utilizando outra operadora
OUTRO CSP	Itens utilizando outro código de seleção de prestadora
OUTROS SERVIÇOS	Outros serviços utilizados
RADIO	Itens utilizando rádio (Ex: Nextel)
SMS	Mensagens enviadas/recebidas
TZ	Itens de tarifa zero (Sem custo)
TZ-MESMA OPERADORA	Tarifa zero com a mesma operadora
TZ-MESMO CSP	Tarifa zero com o mesmo código de seleção de prestadora
VOZ	Ligações do tipo voz
VOZ-RECEBIDA	Ligações do tipo voz que foram recebidas

Deve-se lembrar que cada linha pode ser encontrada em várias seções diferentes da fatura, ou seja, pode possuir mais que um pai, e também pode possuir várias classes associadas à sua classe original. Um exemplo deste tipo de etiquetagem pode ser visualizado na Figura 4.3.

	Linha	Classe Original
1	Ligações com o Código # - Embratel	CONSUMO-VOZ-LD, CONSUMO-VOZ-INTERNACIONAL
2	Ligações com o Código # - Telefônica	CONSUMO-VOZ-LD, CONSUMO-VOZ-INTERNACIONAL
3	Ligações para celulares Claro	CONSUMO-VOZ-LOCAL
4	Ligações recebidas a cobrar	CONSUMO-VOZ-LOCAL, CONSUMO-VOZ-LD

Figura 4.3: Exemplo de etiquetagem utilizando a classificação de forma hierárquica em linhas não folha

Como pode ser observado na Figura 4.3, alguns itens possuem mais de uma classe origi-

nal associada, essas classes são separadas por vírgulas para melhor identificação nas próximas etapas. Deve-se atentar de que as classes apresentadas na figura não são Folhas, ou seja, suas classes não identificam que tipo de item deve ser extraído. Um exemplo da classificação de forma hierárquica para linhas do tipo Folha, podem ser visualizadas na Figura 4.4.

	Linha	Classe Original
1	Pacote Ilimitado Internet #GB - de #/#/# a #/#/# #,#	PACOTE,VALOR-COBRADO
2	#/# #:#:# São Paulo-Promissao #-#-# #:#:# #:#:# #,# #,#	DATA,HORA,NUMERO,DURACAO-EFETIVA,DURACAO,VALOR-TOTAL,VALOR-COBRADO
3	#/# Diaria de Internet Europa Holanda /T-Mobile #,# #,#	DATA,VALOR-TOTAL,VALOR-COBRADO
4	Ligações para telefones fixos #.#,# #,# #,# #,# - #.#,#	CONSUMO,VALOR
5	Chile /Entel PCS # # #,# #,# #,#	QUANTIDADE,TARIFA,VALOR-TOTAL,VALOR-COBRADO
6	Gestor Online - Controle Completo #,#	PACOTE,VALOR

Figura 4.4: Exemplo de etiquetagem utilizando a classificação de forma hierárquica em linhas folha

Na Figura 4.4 é possível observar que a classe original apresenta um formato identificando quais os itens existentes na linha e quais devem ser extraídos para que seja possível preencher o formulário de saída. Após a elaboração do formulário de saída e a etiquetagem do corpus de faturas pelos dois métodos, é possível avançar para a etapa de Pré-Processamento.

4.3 Pré-Processamento

O *corpus* utilizado contém faturas que foram geradas com variações em sua estrutura, como a ordem de algumas seções aparecerem inversamente, a exemplo da seção de Notas Fiscais, que podem aparecer antes, durante ou depois de um detalhamento de número. Nesta etapa, os dados normalmente dispostos em formato inadequado, são preparados para que possam ser utilizados pelo algoritmos de extração de padrões. Diversas tarefas são executadas nesta etapa, como limpeza, transformação e redução de dados.

- Transformação: Com a finalidade de superar quaisquer limitações presentes nos algoritmos de extração de padrões, pode ser necessário modificar a representação dos dados. Como a correção de erros provenientes da leitura das faturas utilizando a biblioteca iText, que faz a conversão de arquivos no formato PDF para o formato TXT.

- Limpeza: Os dados coletados podem apresentar diversos problemas, entre eles instâncias com valores desconhecidos, incorretos ou inválidos, sendo, portanto, importante a realização de uma limpeza nestes dados.

- Redução: Em função de limitações de memória e tempo de processamento é necessário aplicar métodos para redução dos dados. Como a criação de uma máscara, para substituir os itens em formato numérico por um caractere "#". Exemplo: (35,14 em #,#). Essa, auxilia na diminuição da quantidade de linhas iguais apresentadas na fatura, e que são armazenadas em um banco de dados como base de conhecimento.

A biblioteca utilizada na maior parte dos casos realiza perfeitamente a conversão. Entretanto em algumas situações podem ocorrer alguns erros, como separar palavras que visualmente deveriam estar na mesma linha.

A ideia de mascarar um conjunto de caracteres, para a tarefa de etiquetagem/extração, surgiu das seguintes observações:

- Devido as linhas de detalhamento serem muito parecidas, isso diminuiria a quantidade de itens praticamente idênticos da base de conhecimento;
- A padronização de caracteres não afetaria o resultado do sistema de extração, ou seja, as informações que deveriam ser extraídas, com ou sem padronização, seriam as mesmas.

4.3.1 Classificador utilizado para limpeza

Com o objetivo de realizar a etapa de limpeza e identificar as linhas relevantes no processo de EI, utilizou-se um classificador bayesiano linear. Como explicado no Capítulo 2 um classificador, independente de seu formato, necessita extrair características, a fim de realizar os cálculos para determinar a qual classe certo objeto pertence.

A extração das características utilizadas neste processo foram baseadas no estudo sobre o formato das faturas telefônicas, que identifica por exemplo quais os itens relevantes, e como identificá-los. O conjunto dessas características pode ser visualizado na Tabela 4.3.

Tabela 4.3: Identificação de características para o classificador de limpeza

Característica	**Descrição da característica**
Títulos	Conter padrão de título (Nota Fiscal, Documento Financeiro, etc.)
Tem-Cifrão	Possuir ao menos um simbolo $
Alfa-Numérico	Conter tanto letras quanto números
Tem-Sequencia-EouF	Mais que dois (E) ou (F) consecutivos
Tem Padrão Lixo	Possuir padrão de lixo como "(continuação)"ou outros lixos, como "Pág #/#"ou sequencia de letras que formam o código de barras

Para o treinamento do algoritmo foram utilizados 1100 documentos, e para testes foram utilizados os 100 documentos restantes, do total do *corpus* disponibilizado pelo engenheiro do conhecimento. Ambos os documentos foram escolhidos com base em uma função de aleatoriedade, implementado no sistema. O resultado médio entre as faturas de teste utilizando o algoritmo bayesiano foi de 96,7% de acerto entre os itens classificados.

A porcentagem média de erros de 3,3%, dá-se em relação ao número de termos etiquetados como lixo presentes no *corpus* etiquetado, o que torna itens que raramente aparecem nas faturas, e que não apresentam características de lixo, como não-relevantes. Isso deve-se ao fato, da utilização de um único classificador linear, que não consegue identificar com precisão a classe resultante de alguns itens, devido a proporção de itens não-relevantes ser maior do que outras.

É importante observar que diante deste resultado, pode-se concluir que a classificação das informações contidas, foram de alta precisão, o que eleva a confiança para a utilização de classificadores no processo de classificação hierárquica. No entanto, no caso deste trabalho, como tratam-se de faturas telefônicas, que apresentam valores monetários que devem ser extraídos corretamente para não comprometer o resultado apresentado ao usuário, é requisito que a taxa de acerto seja de 100%. Portanto, optou-se por utilizar a associação direta com a base de conhecimento já corretamente etiquetada, o que impossibilita a propagação de erros para a etapa de Extração de Padrões.

4.4 Extração de Padrões

Com a representação dos documentos já em um formato adequado e limpo é possível aplicar o método para a extração de padrões úteis e interessantes presentes nos documentos, de forma que o conhecimento extraído atenda aos objetivos e requisitos do sistema e descritos na etapa de Identificação do Problema.

Com a base de conhecimento já etiquetada, apresentando informações de classes que podem ser usadas para a construção da árvore e a fragmentação das folhas, este trabalho fez uso da técnica de classificação hierárquica por nó pai.

Essa técnica, explicada no Capítulo 2, tem como objetivo criar um classificador linear para cada nó pai encontrado na árvore, identificando sua classe atual. Ao total, foram construídos 32 classificadores bayesianos, que obtiveram individualmente uma taxa de 100% de acerto na

classificação dos itens presentes nos documentos. Suas descrições, características, bem como algumas expressões regulares usadas para identificá-las, podem ser visualizados nas tabelas do Apêndice A.

4.5 Considerações Finais

Este capítulo apresentou as etapas utilizadas para a construção do modelo abordado, suas definições, características necessárias, tanto para a limpeza quanto para a distinção das instâncias dos documentos.

Apresentou o formulário de saída que será utilizado na etapa de experimentos, para armazenar as informações relevantes extraídas pelo sistema. Abordou-se também, as regras de etiquetagem aplicadas para a definição das classes utilizadas na classificação.

Atenta-se à dificuldade da aplicação de uma classificação plana, utilizando apenas um classificador para todo o documento, devido a grande quantidade de instâncias distintas encontradas, e que possuem características semelhantes, o que pode fazer com que o classificador retorne respostas confusas. Por esse motivo é utilizada e abordada a técnica de classificação hierárquica, que separa as instâncias encontradas em determinados classificadores, melhorando seus resultados.

Capítulo 5

Experimentos e Resultados

Este capítulo tem como objetivo descrever como foram elaborados os experimentos e os dados utilizados, apresentando as etapas seguidas para a realização do processamento dos documentos utilizados. Além disso, é apresentada uma análise dos resultados obtidos com a aplicação da técnica de aprendizado de máquina utilizando classificadores hierárquicos, comparando os resultados obtidos sob diferentes conjuntos de treinamento.

5.1 Obtenção e distribuição dos dados

Os dados utilizados nos experimentos deste trabalho foram obtidos através da empresa Orb it Sistemas como já apresentado no capítulo anterior, onde constam um total de 1200 faturas telefônicas. Os experimentos foram realizados inicialmente utilizando um composto de 1195 faturas para treinamento e 5 faturas para testes, as quais foram selecionadas aleatoriamente entre o corpus total disponibilizado.

O critério de avaliação do sistema de extração foi baseado em duas variáveis: $NRevelantes$ expressa o número de informações relevantes que devem ser extraídas da fatura. Essas informações são identificadas previamente com a ajuda do especialista no domínio, que verifica manualmente a quantidade de informações presente nos documentos; e $NCorretos$ se refere ao número de informações totais extraídas corretamente e armazenadas no formulário de saída. A partir deste critério é possível avaliar a precisão do sistema. Em EI, precisão se refere a relação percentual entre a quantidade de informações que devem ser extraídas corretamente e a quantidade de informações totais que foram extraídas dos documentos, e é expressa pela Equação 5.1.

$$preciso = \frac{NRevelantes}{NCorretos} \tag{5.1}$$

Tabela 5.1: Informações das faturas telefônicas utilizadas para testes

Nome Documento	***N* páginas**	***N* de linhas**	**Valor Documento**
Fatura-Teste 1	23	2122	1.297,96 R$
Fatura-Teste 2	13	940	342,67 R$
Fatura-Teste 3	136	8820	10.732,63 R$
Fatura-Teste 4	128	8865	6.472,20 R$
Fatura-Teste 5	72	4991	5.183,61 R$

5.2 Experimentos

Esta seção apresenta uma avaliação geral dos resultados, abordando inicialmente o processo de análise visual dos dados e classificação por meio de consultas no PostgreSQL, SGBD utilizado para armazenar as informações extraídas das faturas. Ao final, é avaliada a solução do processo de extração de informação no escopo das atividades desenvolvidas neste trabalho.

5.2.1 Experimento 1

O primeiro experimento fez uso de 1195 documentos para treinamento, e 5 para testes. Com o objetivo de avaliar a precisão final do sistema de EI utilizando a classificação hierárquica, fazendo uso de uma quantidade significativamente grande de documentos para treinamento.

As Tabelas 5.2, 5.3, 5.4, 5.5, 5.6 e 5.7 mostram a precisão média por informações obtidas pelo sistema, bem como informações relacionadas a cada uma das cinco faturas testadas. Pode-se observar que o resultado obtido em ambos os casos foi de 100% na precisão da extração, utilizando o conjunto de características já citado.

Tabela 5.2: Informações de itens extraídos e quantidade de lixos presentes nas faturas

Nome Documento	***N* linhas**	***N* Itens Extraídos**	**Itens Lixo**	**Valor Documento**
Fatura-Teste 1	2122	1144	978	1.297,96 R$
Fatura-Teste 2	940	478	462	342,67 R$
Fatura-Teste 3	8820	4239	4581	10.732,63 R$
Fatura-Teste 4	8865	5054	3811	6.472,20 R$
Fatura-Teste 5	4991	2214	2214	5.183,61 R$

Observa-se que a quantidade de lixo presente em uma fatura é relativamente grande em comparação ao total de itens relevantes que devem ser extraídos, a fim de preencher o formulário de saída. Esse resultado comprova que a etapa de limpeza de itens não relevantes para a extração é de extrema importância. Por minimizar o custo de processamento e tempo necessários à realização do experimento, uma vez que a quantidade de entradas e comparações necessárias nos classificadores é diminuída.

As Tabelas 5.3, 5.4, 5.5, 5.6, 5.7 apresentam os resultados referentes as informações contidas nas faturas em comparação com os resultados obtidos por meio de consultas no SGBD Postgres. Destaca-se o bom desempenho da estratégia proposta - Classificadores Hierarquicos, utilizando a Classificação por nó pai, que conseguiu obter um resultado de 100% de extração no teste com as cinco faturas.

Tabela 5.3: Documento 1 - Precisão do sistema em relação aos itens extraídos

	Quantidade	**Total itens**	**Itens Extraídos**	**Total na Fatura**	**Total Extraído**	**Precisão**
Nota Fiscal	2	24	24	1185,31	1185,31	100%
Doc-Financeiro	1	1	1	112,65	112,65	100%
Ligações Locais	8	714	714	64,30	64,30	100%
Longa Distância	5	248	248	120,92	120,92	100%
Internacionais	2	7	7	23,70	23,70	100%
Serviços	2	17	17	3,66	3,66	100%

Tabela 5.4: Documento 2 - Precisão do sistema em relação aos itens extraídos

	Quantidade	**Total itens**	**Itens Extraídos**	**Total na Fatura**	**Total Extraído**	**Precisão**
Nota Fiscal	2	18	18	326,59	326,59	100%
Doc-Financeiro	1	1	1	16,08	16,08	100%
Ligações Locais	3	353	353	41,79	41,79	100%
Longa Distância	2	62	62	0,00	0,00	100%
Serviços	2	6	6	0,30	0,30	100%

Tabela 5.5: Documento 3 - Precisão do sistema em relação aos itens extraídos

	Quantidade	**Total itens**	**Itens Extraídos**	**Total na Fatura**	**Total Extraído**	**Precisão**
Nota Fiscal	4	40	40	8423,45	8423,45	100%
Doc-Financeiro	1	4	4	2309,28	2309,28	100%
Ligações Locais	3	2311	2311	311,03	3011,03	100%
Longa Distância	2	670	670	375,31	375,31	100%
Serviços	2	29	29	5,52	5,52	100%

Tabela 5.6: Documento 4 - Precisão do sistema em relação aos itens extraídos

	Quantidade	Total itens	Itens Extraídos	Total na Fatura	Total Extraído	Precisão
Nota Fiscal	7	20	20	4765,09	4765,09	100%
Doc-Financeiro	1	3	3	1707,11	1707,11	100%
Ligações Locais	57	2207	2207	0,00	0,00	100%
Longa Distância	2	2308	2308	0,00	0,00	100%
Serviços	10	16	16	0,00	0,00	100%

Tabela 5.7: Documento 5 - Precisão do sistema em relação aos itens extraídos

	Quantidade	Total itens	Itens Extraídos	Total na Fatura	Total Extraído	Precisão
Nota Fiscal	3	25	25	4723,52	4723,52	100%
Doc-Financeiro	1	2	2	460,10	460,10	100%
Ligações Locais	22	1925	1925	0,00	0,00	100%
Longa Distância	18	369	369	0,00	0,00	100%
Serviços	10	28	28	0,00	0,00	100%

A Tabela 5.6 mostra uma fatura que se difere das outras visto que apresenta um total de sete notas fiscais, apresentando seções que não possuem itens internos, e portanto, não foram extraídos. Isso ocorre muitas vezes por um erro na geração da fatura por parte da operadora. Essas seções não interferem no resultado do sistema, já que como não contém itens internos, e não possuem um valor monetário a ser extraído.

A Tabela 5.7 indica que o valor total extraído da fatura, somando o documento financeiro e as notas fiscais é de R$ 5183,62, sendo seu valor real como apresentado na Tabela 5.1 R$ 5183,61. Essa discrepância pode ser identificada pelo fato da fatura conter um Crédito Anterior no valor de R$ -0,01, que não está presente em uma seção relevante, e por esse motivo não foi extraído.

5.2.2 Experimento 2

Com o intuito de demonstrar a precisão do sistema utilizando um conjunto de treinamento menor que o apresentado no primeiro experimento, essa segunda etapa de testes utiliza uma variação na quantidade de documentos usados na etapa de treinamento do sistema, sendo respectivamente 100, 200, 400, 800 e 1000 faturas, aplicadas para cada um dos 5 documentos utilizados nos testes.

As Tabelas 5.8, 5.9, 5.10, 5.11 e 5.12 mostram a precisão na extração de informações nas mesmas cinco faturas utilizadas no teste anterior.

Tabela 5.8: Precisão do sistema utilizando 100 documentos para treinamento

	Tempo Treinamento (ms)	Tempo Teste (ms)	Instâncias Treino	Instâncias Teste	Precisão
Documento-1	30927	3580	1358057	2122	100%
Documento-2	30856	2264	1358057	940	100%
Documento-3	30915	5336	1358057	8820	100%
Documento-4	30924	5325	1358057	8865	100%
Documento-5	30855	3899	1358057	4991	100%

Tabela 5.9: Precisão do sistema utilizando 200 documentos para treinamento

	Tempo Treinamento (ms)	Tempo Teste (ms)	Instâncias Treino	Instâncias Teste	Precisão
Documento-1	60549	3415	3456758	2122	100%
Documento-2	60536	2235	3456758	940	100%
Documento-3	60485	5332	3456758	8820	100%
Documento-4	60554	5315	3456758	8865	100%
Documento-5	60549	3844	3456758	4991	100%

Tabela 5.10: Precisão do sistema utilizando 400 documentos para treinamento

	Tempo Treinamento (ms)	Tempo Teste (ms)	Instâncias Treino	Instâncias Teste	Precisão
Documento-1	140771	3578	8284116	2122	100%
Documento-2	140689	2224	8284116	940	100%
Documento-3	141068	5332	8284116	8820	100%
Documento-4	141005	5332	8284116	8865	100%
Documento-5	140884	3890	8284116	4991	100%

Tabela 5.11: Precisão do sistema utilizando 800 documentos para treinamento

	Tempo Treinamento (ms)	Tempo Teste (ms)	Instâncias Treino	Instâncias Teste	Precisão
Documento-1	223966	3486	13314227	2122	100%
Documento-2	224056	2243	13314227	940	100%
Documento-3	224104	5343	13314227	8820	100%
Documento-4	223969	5377	13314227	8865	100%
Documento-5	223978	3868	13314227	4991	100%

Tabela 5.12: Precisão do sistema utilizando 1000 documentos para treinamento

	Tempo Treinamento (ms)	Tempo Teste (ms)	Instâncias Treino	Instâncias Teste	Precisão
Documento-1	358112	3512	22537473	2122	100%
Documento-2	357865	2239	22537473	940	100%
Documento-3	358115	5320	22537473	8820	100%
Documento-4	357854	5420	22537473	8865	100%
Documento-5	358012	3888	22537473	4991	100%

Pode-se observar que a precisão do sistema, mesmo utilizando um conjunto de treinamento reduzido, mantêm-se idêntica. Porém, como esperado, à medida que a quantidade de documentos utilizados aumenta, o tempo de execução cresce, pelo fato de ser necessária a leitura de cada um dos elementos encontrados nos documentos, pelo sistema extrator de informação.

5.3 Considerações Finais

Este capítulo apresentou os resultados dos testes realizados com o sistema desenvolvido neste trabalho. Realizou-se cinco testes utilizando faturas distintas, com o intuito de avaliar o comportamento do sistema sob diferentes condições de teste.

O experimento mostrou um desempenho significativo do sistema utilizando a mesma quantidade de documento para treinamento em ambos os cinco casos de teste, totalizando um resultado de 100% na extração das informações.

O segundo experimento realizado mostrou a precisão obtida pelo sistema com diferentes tamanhos do conjunto de treinamento, observando que a taxa de acerto do sistema se manteve alta, mesmo utilizando uma quantidade relativamente menor de faturas de treinamento.

Percebe-se que devido a quantidade de elementos encontrados nas faturas e a leitura de todos eles utilizando a biblioteca iText, o tempo de execução aumenta linearmente à medida que aumenta-se a quantidade de arquivos usados para treinamento.

Os resultados obtidos pelo sistema não puderam ser comparados com outros trabalhos de extração de informação, por serem inexistentes trabalhos relacionados com essa área.

Capítulo 6

Considerações Finais

O objetivo deste trabalho foi a construção de um sistema para realizar a extração de informação de faturas de telefonia móvel corporativa por meio de uma abordagem baseada em aprendizagem de máquina. Para isso, foi realizado um estudo das principais técnicas utilizadas pelos sistemas *wrappers* para extração de informação.

Este estudo possibilitou a criação de um sistema que utiliza uma abordagem de classificação hierárquica utilizando para isso classificadores bayesianos. Como mostrado, esse sistema conseguiu obter uma classificação eficaz de 100% para a tarefa proposta, porém utilizando a associação direta com a base de conhecimento na etapa de pré-processamento, mais precisamente na etapa de limpeza.

6.1 Contribuições

A principal contribuição deste trabalho foi a utilização de uma abordagem de classificação hierárquica para a extração de informação em faturas de telefonia móvel corporativa por meio de um *wrapper* que utilizou um conjunto de classificadores bayesianos. Foram realizados diversos experimentos que mostraram que, no domínio das faturas telefônicas, esta abordagem consegue extrair as informações especificadas pelo usuário com precisão.

Foi desenvolvido um sistema para EI em faturas telefônicas capaz de lidar com variados formatos de estruturação, e que pode vir a ser utilizado na prática em empresas que realizam a gestão dos serviços de telecomunicações, a fim de auxiliar no processo de tomada de decisão. Além disso, foi realizado um estudo das técnicas de extração de informação através de classificadores, aplicando-se no experimento deste trabalho. Foi abordado ainda que para esse tipo

de trabalho, por basear-se em auxiliar no processo de gestão de telecomunicações, não se pode tolerar erros de extração e/ou representação, como no processo de limpeza, que não gerou uma taxa de acerto condizente em seu classificador.

6.2 Dificuldades

Uma das grandes dificuldades encontradas neste trabalho, foi o estudo e elaboração das características utilizadas pelos classificadores a fim de determinar os itens relevantes e não relevantes dos documentos. Isso se deve ao fato de uma fatura telefônica possuir uma estrutura não regular e linhas bastante similares em seus tópicos internos.

Primeiramente, tentou-se utilizar apenas um classificador linear para extrair as informações de toda a fatura. Entretanto, pelo mesmo não ter apresentado resultados satisfatórios, aplicou-se a técnica de classificação hierárquica, facilitando a elaboração de características e melhorando o resultado final do sistema.

6.3 Trabalhos Futuros

O trabalho desenvolvido obteve resultados bastante satisfatórios para a tarefa de EI em faturas telefônicas corporativas. Porém, são possíveis melhorias que podem não apenas aprimorar os resultados parciais do processo de extração atual, como facilitar o uso e diminuir o tempo da atividade, principalmente na etapa de Extração de Padrões. Sugestões relacionadas com esses procedimentos são apresentadas a seguir.

- **Aprimoramento da etapa de limpeza utilizando HMM**

  Os Modelos de Markov Escondidos (HMM, do inglês *Hidden Markov Models*), segundo [Rabiner e Juang 1986] são capazes de explorar naturalmente a ocorrência dos padrões em sequência no texto de entrada para classificá-los de uma só vez, por meio de uma classificação que maximize a probabilidade de acerto para todo o conjunto de padrões.

  A ideia básica, é realizar a combinação da etapa de limpeza gerando uma saída inicial para o sistema, utilizando as técnicas de classificação, e depois refiná-las por meio de um HMM. Esse refinamento tem como objetivo, o melhoramento do resultado de limpeza, sem que seja necessária a consulta direta à base de conhecimento.

- **Aplicação da técnica RBC na Extração de Informação**

  O Raciocínio Baseado em Casos (RBC), segundo [Abel 1996] é uma técnica que busca resolver novos problemas adaptando soluções utilizadas para resolver problemas anteriores. Os sistemas RBC representam um dos grandes sucessos das pesquisas da inteligência artificial, devido ao fato de apresentarem grande abrangência de aplicação. Essa técnica, ainda segundo [Abel 1996] se baseia na premissa de que seres humanos utilizam um raciocínio analógico ou experimental para aprender a resolver problemas complexos. Neste sentido, com o objetivo de melhorar os resultados obtidos na etapa de limpeza, e os resultados da aplicação desta técnica podem ser comparados com os da extração realizada neste trabalho.

Apêndice A

Construção dos Classificadores

Classificador Raiz

Este classificador tem como objetivo identificar as classes filhas do classificador Raiz: DETALHAMENTO, NOTA FISCAL e DOC-FINANCEIRO. Possui três características, como mostra a Tabela A.1.

Tabela A.1: Características Classificador Raiz

Características	Retorno	Descrição
temTextoDetalhamento	booleano	Tem texto "detalhamento"
temNotaFiscal	booleano	Tem texto "nota fiscal"
temDocumentoFinanceiro	booleano	Tem texto "documento financeiro"

Classificador Detalhamento

Este classificador tem como objetivo identificar as classes filhas do classificador Detalhamento: CONSUMO-EXTERIOR, CONSUMO, CONSUMO-VOZ-INTERNACIONAL, CONSUMO-VOZ-LOCAL, CONSUMO-VOZ-MESMA OPERADORA, MENSALIDADES E PACOTES, CONSUMO-VOZ-LD, CONSUMO-VOZ-SERVIÇO e CONSUMO-VOZ. Possui nove características, como mostra a Tabela A.2.

Tabela A.2: Características Classificador Detalhamento

Características	Retorno	Descrição
temTextoInterurbano	booleano	Tem texto "interurbanas"
temTextoLocal	booleano	Tem texto "locais"
temTextoMensalidades	booleano	Tem texto "mensalidades"
temTextoInternacionais	booleano	Tem texto "internacionais"
temTextoVideo	booleano	Tem texto "vídeo"
temTextoEspeciais	booleano	Tem texto "especiais"
temTextoOnNet	booleano	Tem texto "on net"
temTextoExterior	booleano	Tem texto "exterior"
temTextoServicos	booleano	Tem texto "serviços"

Classificador Mensalidade

Este classificador tem como objetivo identificar as classes filhas do classificador Mensalidades. Diferente dos classificadores intermediários mostrados acima, este classificador determina a classe das linhas FOLHA, com: PLANO-VALOR, DESCONTO-VALOR, PACOTE-VALOR, PACOTE, VALOR-COBRADO, GERAL, VALOR. Possui quatro características, como mostra a Tabela A.3.

Tabela A.3: Características Classificador Mensalidade

Características	Retorno	Descrição
temTextoMulta	booleano	Tem texto "multa"
temTextoPacote	booleano	Tem texto "pacote"ou "pct"ou "bônus"
temTextoPlano	booleano	Tem texto "plano"ou "assinatura"ou "pl."ou "consumo"
temTextoDesconto	booleano	Tem texto "desconto"

Classificador Consumo no Exterior

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumo no Exterior: VOZ, DADOS, SMS, VOZ-RECEBIDA. Possui quatro características, como mostra a Tabela A.4.

Tabela A.4: Características Classificador Consumo Exterior

Características	Retorno	Descrição
temTextoFeitas	booleano	Tem texto "feitas"
temTextoRecebidas	booleano	Tem texto "recebidas"
temTextoDados	booleano	Tem texto "dados"
temTextoTorpedos	booleano	Tem texto "torpedos"

Classificador Consumo

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumos: DADOS, SMS, OUTROS SERVIÇOS. Possui três características, como mostra a Tabela A.5.

Tabela A.5: Características Classificador Consumo

Características	Retorno	Descrição
temTextoDados	booleano	Tem texto "dados"
temTextoTorpedos	booleano	Tem texto "torpedos"
temTextoServicos	booleano	Tem texto "serviços"

Classificador Consumo de Voz Internacional

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumos de Voz Internacional: MESMO CSP. Possui apenas uma característica, como mostra a Tabela A.6.

Tabela A.6: Características Classificador Consumo Voz Internacional

Características	Retorno	Descrição
temTextoInternacionais	booleano	Tem texto "internacionais"

Classificador Consumo de Voz Local

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumos de Voz Local: OUTRA OPERADORA, TZ-MESMA OPERADORA, DESLOCAMENTO, TZ, MESMA OPERADORA, RADIO, FIXO, A COBRAR-RECEBIDA. Possui sete características, como mostra a Tabela A.7.

Tabela A.7: Características Classificador Consumo Voz Local

Características	Retorno	Descrição
temTextoOutras	booleano	Tem texto "outras"
temTextoTarifaZero	booleano	Tem texto "tarifa zero"
temTextoForaCobertura	booleano	Tem texto "fora da cobertura"
temTextoClaro	booleano	Tem texto "claro"
temTextoRadio	booleano	Tem texto "rádio"
temTextoFixo	booleano	Tem texto "fixo"
temTextoAcobrar	booleano	Tem texto "cobrar"

Classificador Consumo de Voz Mesma Operadora

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumos de Voz Mesma Operadora. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO. Possui duas características, como mostra a Tabela A.8.

Tabela A.8: Características Classificador de Consumos de Voz Mesma Operadora

Características	Retorno	Descrição
temTextoONnet	booleano	Tem texto "ONnet"
temTextoTarifaZero	booleano	Expressão Regular Regex "(w+(\d+)?)\s+((\d+.)?\d+,\d+\s+){2}(\d+.)?\d+,\d+$"

Classificador Consumo de Voz Longa Distância

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumos de Voz Longa Distância: MESMO CSP, DESLOCAMENTO-RECEBIDA, TZ-MESMO CSP, EXTRAGRUPO-MESMA OPERADORA, OUTRO CSP, RADIO, INTRAGRUPO-MESMA OPERADORA. Possui oito características, como mostra a Tabela A.9.

Tabela A.9: Características Classificador Consumo de Voz Longa Distância

Características	Retorno	Descrição
temTextoAcobrar	booleano	Tem texto "a cobrar"
temTextoIntraGrupo	booleano	Tem texto "intra-grupo"
temTextoRadio	booleano	Tem texto "rádio"
temTextoCodigo	booleano	Tem texto "código"
temTextoEmbratel	booleano	Tem texto "embratel"
temTextoViagem	booleano	Tem texto "viagem"
temTextoTarifaZero	booleano	Tem texto "tarifa zero"
temTextoIntraRede	booleano	Tem texto "intra-rede"

Classificador Consumo de Voz Serviços

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumos de Voz Serviços. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO. Possui duas características, como mostra a Tabela A.10.

Tabela A.10: Características Classificador Consumo de Voz Serviços

Características	Retorno	Descrição
temPadraoNumero	booleano	Expressão Regular Regex "\s+(\d+-\d+-\d+|\d+|*\d+)\s+"
temPadraoSeguidoDuracao	booleano	Expressão Regular Regex "\s+(\d+-\d+-\d+|\d+|*\d+)\s+\d+:\d+:\d+"

Classificador Consumo de Voz

Este classificador tem como objetivo identificar as classes filhas do classificador de Consumos de Voz: OUTRA OPERADORA, MESMA OPERADORA, LD. Possui três características, como mostra a Tabela A.11.

Tabela A.11: Características Classificador Consumo de Voz

Características	Retorno	Descrição
temTextoOutras	booleano	Tem texto "outras"
temTextoClaro	booleano	Tem texto "claro"
temTextoInterurbanas	booleano	Tem texto "interurbanas"

Classificador de Voz

Este classificador tem como objetivo identificar as classes filhas do classificador de Voz. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO,DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, DURACAO, VALOR-TOTAL, VALOR-COBRADO. Possui quatro características, como mostra a Tabela A.12.

Tabela A.12: Características Classificador de Voz

Características	Retorno	Descrição
temTextoNacInter	booleano	Tem texto "Nacional"ou "Internacional"
temDoisValoresSequencia	booleano	Expressão Regular Regex "\s+(\d+.)?\d+,\d+\s+(\d+.)?\d+,\d+$"
temPadraoNumero	booleano	Expressão Regular Regex "\s+\d+\s+"
temPadraoSeguidoDuracao	booleano	Expressão Regular Regex "\s+(\d+-\d+-\d+|\d+|*\d+)\s+\d+:\d+:\d+"

Classificador de Dados

Este classificador tem como objetivo identificar as classes filhas do classificador de Dados. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: MB-UTILIZADOS, VALOR-COBRADO ou DATA, VALOR-TOTAL, VALOR-COBRADO ou DATA, MB-UTILIZADOS, VALOR-TOTAL, VALOR-COBRADO ou DATA, VALOR-TOTAL, VALOR-COBRADO ou MB-UTILIZADOS, VALOR-COBRADO. Possui quatro características, como mostra a Tabela A.13.

Tabela A.13: Características Classificador de Dados

Características	Retorno	Descrição
temPadraoData	booleano	Expressão Regular Regex "^\d+\/\d+"
temDoisValoresSequencia	booleano	Expressão Regular Regex "(\w+\d+?l\s+\d+)\s+(\d+\.)?\d+,\d+\s+(\d+\.)?\d+,\d+$"
temTresValoresSequencia	booleano	Expressão Regular Regex "(\w+(\d+)?)\s+((\d+.)?\d+,\d+\s+){2}(\d+.)?\d+,\d+$"
temQuatroValoresSequencia	booleano	Expressão Regular Regex "(\w+(\d+)?)\s+((\d+.)?\d+,\d+\s+){3}(\d+.)?\d+,\d+$"

Classificador de SMS

Este classificador tem como objetivo identificar as classes filhas do classificador de SMS. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, QUANTIDADE, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, QUANTIDADE, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou QUANTIDADE, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, TARIFA, VALOR-TOTAL, VALOR-COBRADO. Possui quatro características, como mostra a Tabela A.14.

Tabela A.14: Características Classificador de SMS

Características	Retorno	Descrição
temPadraoData	booleano	Expressão Regular Regex "^\d+\/\d+"
temPadraoHora	booleano	Expressão Regular Regex "\s+(\d+:\d+:\d+)\s+"
temPadraoQuantidade	booleano	Expressão Regular Regex "(\s+\d+\s+)l(\w+\s+(((\d+\.)?\d+,\d+)\s+){3}(\d+\.)?\d+,\d+)"
temPadraoNumero	booleano	Expressão Regular Regex "\s+(\d+-\d+-\d+)\s+"

Classificador de Voz Recebida

Este classificador tem como objetivo identificar as classes filhas do classificador de Voz Recebida. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, DURACAO, VALOR-TOTAL, VALOR-COBRADO. Possui duas características, como mostra a Tabela A.15.

Tabela A.15: Características Classificador de Voz Recebida

Características	Retorno	Descrição
temPadraoNumero	booleano	Expressão Regular Regex "\s+\d+\s+"
temPadraoSeguidoDuracao	booleano	Expressão Regular Regex "\s+\d+\s+\d+:\d+:\d+"

Classificador de Outros Serviços

Este classificador tem como objetivo identificar as classes filhas do classificador de Outros Serviços. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: QUANTIDADE, TARIFA, VALOR-TOTAL, VALOR-COBRADO. Possui duas características, como mostra a Tabela A.16.

Tabela A.16: Características Classificador de Outros Serviços

Características	Retorno	Descrição
temPadraoQuantidade	booleano	Expressão Regular Regex "\w+\s+\d+\s+"
temTresValoresSequencia	booleano	Expressão Regular Regex "(((\d+\.)?\d+,\d+)\s+){2}((\d+\.)?\d+,\d+)"

Classificador de Mesmo CSP

Este classificador tem como objetivo identificar as classes filhas do classificador de Mesmo CSP. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO. Possui duas características, como mostra a Tabela A.17.

Tabela A.17: Características Classificador de Mesmo CSP

Características	Retorno	Descrição
temPadraoNumero	booleano	Expressão Regular Regex "\s+(\d+-\d+-\d+|\d+)\s+"
temDoisPadroesDuracao	booleano	Expressão Regular Regex "(\s+\d+:\d+:\d+){2}\s+"

Classificador de Outro CSP

Este classificador tem como objetivo identificar as classes filhas do classificador de Mesmo CSP. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO. Possui três características, como mostra a Tabela A.18.

Tabela A.18: Características Classificador de Outro CSP

Características	Retorno	Descrição
temPadraoNumero	booleano	Expressão Regular Regex "\s+(\d+-\d+-\d+|\d+)\s+"
temUmPadraoDuracao	booleano	Expressão Regular Regex "\s+\d+:\d+:\d+\s+"
temDoisPadroesDuracao	booleano	Expressão Regular Regex "\s+((\d+:\d+:\d+)\s+){2}"

Classificador de Outra Operadora

Este classificador tem como objetivo identificar as classes filhas do classificador de Outra Operadora. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO. Possui duas características, como mostra a Tabela A.19.

Tabela A.19: Características Classificador de Outra Operadora

Características	Retorno	Descrição
temPadraoNumero	booleano	Expressão Regular Regex "\s+(\d+-\d+-\d+|\d+)\s+"
temDoisPadroesDuracao	booleano	Expressão Regular Regex "\s+((\d+:\d+:\d+)\s+){2}"

Classificador de Tarifa Zero Mesma Operadora

Este classificador tem como objetivo identificar as classes filhas do classificador de Tarifa Zero Mesma Operadora. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO. Possui duas características, como mostra a Tabela A.20.

Tabela A.20: Características Classificador Tarifa Zero Mesma Operadora

Características	Retorno	Descrição
temUmPadraoDuracao	booleano	Expressão Regular Regex "(\s+\d+:\d+:\d+\s+)"
temDoisPadroesDuracao	booleano	Expressão Regular Regex "\s+((\d+:\d+:\d+)\s+){2}"

Classificador de Deslocamento

Este classificador tem como objetivo identificar as classes filhas do classificador de Deslocamento. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de Tarifa Zero

Este classificador tem como objetivo identificar as classes filhas do classificador de Tarifa Zero. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de Mesma Operadora

Este classificador tem como objetivo identificar as classes filhas do classificador de Mesma Operadora. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de Rádio

Este classificador tem como objetivo identificar as classes filhas do classificador de Rádio. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO. Possui três características, como mostra a Tabela A.21.

Tabela A.21: Características Classificador de Rádio

Características	Retorno	Descrição
temTresValoresSequencia	booleano	Expressão Regular Regex "(\w+(\d+)?)\s+((\d+.)?\d+,\d+\s+){2}(\d+.)?\d+,\d+$"
temUmPadraoDuracao	booleano	Expressão Regular Regex "\s+\d+:\d+:\d+\s+"
temDoisPadroesDuracao	booleano	Expressão Regular Regex "\s+((\d+:\d+:\d+)\s+){2}"

Classificador de Fixo

Este classificador tem como objetivo identificar as classes filhas do classificador de Fixo. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de A Cobrar Recebida

Este classificador tem como objetivo identificar as classes filhas do classificador de Outra Operadora. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO, TARIFA, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Rádio.

Classificador de Deslocamento Recebida

Este classificador tem como objetivo identificar as classes filhas do classificador de Deslocamento Recebida. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de Tarifa Zero Mesmo CSP

Este classificador tem como objetivo identificar as classes filhas do classificador de Tarifa Zero Mesmo CSP. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de Extragrupo Mesma Operadora

Este classificador tem como objetivo identificar as classes filhas do classificador de Extragrupo Mesma Operadora. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de Intragrupo Mesma Operadora

Este classificador tem como objetivo identificar as classes filhas do classificador de Intragrupo Mesma Operadora. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DATA, HORA, NUMERO, DURACAO, VALOR-TOTAL, VALOR-COBRADO ou DATA, HORA, NUMERO, DURACAO-EFETIVA, DURACAO, VALOR-TOTAL, VALOR-COBRADO.

As características usadas por esse classificador são as mesmas usadas pelo classificador de Tarifa Zero Mesma Operadora.

Classificador de Nota Fiscal

Este classificador tem como objetivo identificar as classes filhas do classificador de Nota Fiscal. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: DESCONTO, VALOR ou PLANO, VALOR ou CONSUMO, VALOR ou PACOTE, VALOR ou GERAL, VALOR. Possui cinco características, como mostra a Tabela A.22.

O mesmo possui cinco características:

Tabela A.22: Características Classificador de Nota Fiscal

Características	Retorno	Descrição
temTextoOutros	booleano	Tem texto "Outros"
temTextoDesconto	booleano	Tem texto "Desconto"
temTextoPlano	booleano	Tem texto "Pl.", "Plano"ou "Assinatura"
temTextoLigacoes	booleano	Tem texto "Ligação", "Chamadas", "Dados", "Interurbanas", "Torpedos"ou "Auxílio"
temTextoPacote	booleano	Tem texto "Compartilhado", "Pct", "Pacote", "Total", "Módulo", "Banda", "Comunidade"ou "Serviço"

Classificador de Documento Financeiro

Este classificador tem como objetivo identificar as classes filhas do classificador de Documento Financeiro. Diferente dos classificadores intermediários, determina a classe das linhas FOLHA com: PLANO,VALOR ou DESCONTO,VALOR ou CONSUMO,VALOR ou PACOTE,VALOR ou GERAL,VALOR.

O mesmo possui quatro características:

Tabela A.23: Características Classificador de Documento Financeiro

Características	**Retorno**	**Descrição**
temTextoAssinatura	booleano	Tem texto "rastreamento"
temTextoDesconto	booleano	Tem texto "desconto"
temTextoLigacoes	booleano	Tem texto "ligações"
temTextoParcelamento	booleano	Tem texto "multa"ou "parcela"

Referências

[Abel 1996]ABEL, M. *Um estudo sobre Raciocínio Baseado em Casos*. Porto Alegre, RS, Brasil, 1996. Disponível em: <http://www.inf.ufrgs.br/bdi/wp-content/uploads/CBR-TI60.pdf>.

[Appelt e Israel 1999]APPELT, D.; ISRAEL, D. Introduction to information extraction. In: . Amsterdam, The Netherlands, The Netherlands: IOS Press, 1999. v. 12, n. 3, p. 161–172. Disponível em: <http://dl.acm.org/citation.cfm?id=1216155.1216161>.

[Burred e Lerch 2003]BURRED, J. J.; LERCH, A. A hierarquical approach to automatic musical genre classification. *In Proceedings of the 6th International Conference on Digital Audio Effects*, p. 8–11, 2003.

[Carvalho 2011]CARVALHO, R. V. Um método evolucionário para classificação hierárquica. UFMG, Belo Horizonte, MG, Brasil, May 2011. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/SLSS-8HTM7W/rafaelvieiracarvalho.pdf?sequence=1>.

[Cavalcante 2009]CAVALCANTE, J. R. R. Gestão de custos em telecom. In: . Rio de Janeiro, RJ, Brasil: E-papers, 2009. (1ªedição). ISBN 978-85-7650-198-5.

[Corrêa, Marcacini e Rezende 2012]CORRêA, G. N.; MARCACINI, R. M.; REZENDE, S. O. Uso da mineração de textos na análise exploratória de artigos científicos. ICMC-USP, São Carlos, SP, Brasil, n. 383, 2012.

[Costa et al. 2007]COSTA, E. P.; LORENA, A. C.; CARVALHO, A. C. P. L. F.; FREITAS, A. A. A review of performance evaluation measures for hierarchical classifiers. In: DRUMMOND, C. et al. (Ed.). *Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop, AAAI Technical Report WS-07-05*. AAAI Press, 2007. p. 182–196. ISBN 978-1-57735-332-4. Disponível em: <http://www.cs.kent.ac.uk/pubs/2007/2611>.

[Cowie e Lehnert 1996]COWIE, J. R.; LEHNERT, W. G. Information extraction. *Commun. ACM*, v. 39, p. 80–91, 1996.

[Dorre, Gerstl e Seiffert 1999]DORRE, J.; GERSTL, P.; SEIFFERT, R. Text mining: Finding nuggets in mountains of textual data. *In KDD '99: Proceedings of Thte Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, p. 398–401, 1999.

[Feldman e Dagan 1995]FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases (kdt). In: *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95*. Montreal, Quebec, Canada: AAAI Press, 1995. p. 112–117.

[Feldman e Hirsh 1997]FELDMAN, R.; HIRSH, H. Explointing background information in knowledge discovery from text. *Journal of Intelligent Information Systems*, v. 9, n. 1, p. 83–97, January 1997.

[Hand, Mannila e Smyth 2001]HAND, D.; MANNILA, H.; SMYTH, P. *Principles of Data Mining*. MIT Press, 2001. (A Bradford book). ISBN 9780262082907. Disponível em: <https://books.google.com.br/books?id=SdZ-bhVhZGYC>.

[Hopcroft e Ullman 1979]HOPCROFT, J. E.; ULLMAN, J. D. *Introduction to Automata Theory, Languages, and Computation*. Boston, MA, USA: Addison-Wesley Publishing Company, 1979.

[Jackson e Moulinier 2002]JACKSON, P.; MOULINIER, I. *Natural Language Processing for Online Applications : Text Retrieval, Extraction, and Categorization*. Amsterdam, Philadelphia: J. Benjamins, 2002. (Natural language processing). ISBN 90-272-4988-1. Disponível em: <http://opac.inria.fr/record=b1098800>.

[Jacobs e Rau 1993]JACOBS, P. S.; RAU, L. F. Innovations in text interpretation. *Artif. Intell.*, Elsevier Science Publishers Ltd., Essex, UK, v. 63, n. 1-2, p. 143–191, out. 1993. ISSN 0004-3702. Disponível em: <http://dx.doi.org/10.1016/0004-3702(93)90016-5>.

[Jubran, Urbano e Risso 1992]JUBRAN, C. C. A. S.; URBANO, H.; RISSO, M. S. Organização tópica na conversação. *In: ILARI, R (Org.). Gramática do Português Falado*, Editora da UNICAMP; FAPESP, Campinas, SP, Brasil, v. 2, p. 357–440, 1992.

[Kushmerick e Thomas 2003]KUSHMERICK, N.; THOMAS, B. Adaptive information extraction: Core technologies for information agents. In: *AgentLink*. Berlin, Heidelberg, Alemanha: Springer, 2003. (Lecture Notes in Computer Science, v. 2586), p. 79–103.

[Lowagie 2010]LOWAGIE, B. *iText in Action*. Greenwich, CT, USA: Manning Publications Co., 2010. ISBN 1935182617, 9781935182610.

[Álvarez 2007]ÁLVAREZ, A. C. Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem. Biblioteca Digital de Teses e Dissertações da USP, São Carlos, SP, Brasil, May 2007. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-21062007-144352/pt-br.php>.

[Menezes 2014]MENEZES, R. A. Construção de classificadores hierárquicos multirrótulo usando evolução diferencial. *Pontifícia Universidade Católica do Paraná*, Curitiba, PR, Brasil, nov. 2014.

[Paes 2012]PAES, B. C. *Novas Estratégias Para Classificação Hierárquica Local Por Nível*. Dissertação (Dissertação de Mestrado) — Universidade Federal Fluminense - UFF, Oct 2012.

[Rabiner e Juang 1986]RABINER, L. R.; JUANG, B. H. An introduction to hidden markov models. *IEEE ASSp Magazine*, 1986.

[Rezende 2003]REZENDE, S. *Sistemas inteligentes: fundamentos e aplicações*. Manole, 2003. ISBN 9788520416839. Disponível em: <https://books.google.com.br/books?id=UsJe_PlbnWcC>.

[Riloff 1999]RILOFF, E. Understanding language understanding. In: RAM, A.; MOORMAN, K. (Ed.). Cambridge, MA, USA: MIT Press, 1999. cap. Information Extraction As a Stepping Stone Toward Story Understanding, p. 435–460. ISBN 0-262-18192-4. Disponível em: <http://dl.acm.org/citation.cfm?id=305135.305152>.

[Rossoni, Pires e Zanotta 2013]ROSSONI, F. da R.; PIRES, V. J.; ZANOTTA, D. Mapeamento urbano por classificação hierárquica semi-automática baseada em objetos. *Anais XVI Simpósio Brasileiro de Sensoriamento Remoto - SBSR*, Foz do Iguaçu, PR, Brasil, aug 2013.

[Secker et al. 2007]SECKER, A.; DAVIES, M. N.; FREITAS, A. A.; TIMMIS, J.; MENDAO, M.; FLOWER, D. R. *An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function*. 2007.

[Silla e Freitas 2011]SILLA, C. N.; FREITAS, A. A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.*, v. 22, p. 31–72, 2011.

[Silva 2004]SILVA, E. F. do Amaral e. *Um Sistema para Extração de Informação em Referências Bibliográficas Baseado em Aprendizado de Máquina*. Dissertação (Dissertação de Mestrado) — Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil, 2004.

[Soderland et al. 1995]SODERLAND, S.; FISHER, D.; ASELTINE, J.; LEHNERT, W. G. Crystal: Inducing a conceptual dictionary. *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, p. 1314–1319, 1995.

[Sun e Lim 2001]SUN, A.; LIM, E.-P. Hierarchical text classification and evaluation. In: *Proceedings of the 2001 IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2001. (ICDM '01), p. 521–528. ISBN 0-7695-1119-8. Disponível em: <http://dl.acm.org/citation.cfm?id=645496.657884>.

[Vens et al. 2008]VENS, C.; STRUYF, J.; SCHIETGAT, L.; DZEROSKI, S.; BLOCKEEL, H. Decision trees for hierarchical multi-label classification. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 73, n. 2, p. 185–214, nov. 2008. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1007/s10994-008-5077-3>.

[Yang e Pedersen 1997]YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. (ICML '97), p. 412–420. ISBN 1-55860-486-3. Disponível em: <http://dl.acm.org/citation.cfm?id=645526.657137>.

[Zambenedetti e Oliveira 2002]ZAMBENEDETTI, C.; OLIVEIRA, J. P. M. de. Extração de informação sobre bases de dados textuais. Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, RS, Brasil, 2002.

[Zhang 2004]ZHANG, H. The optimality of naive bayes. In: BARR, V.; MARKOV, Z. (Ed.). *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. Miami Beach, Florida, USA: AAAI Press, 2004.