DOI: 10.5748/9788599693148-15CONTECSI/PS-5752

Guilherme Felipe Zabot, 0000-0002-5550-2644, (Universidade de São Paulo, São Paulo, Brasil) – zabot.gui@gmail.com

Gustavo Rezende Krüger, 0000-0003-1914-1156, (Orb it Sistemas, Paraná, Brasil) - gustavorekr@gmail.com

Marcio Seiji Oyamada, 0000-0002-6354-8917, (Universidade Estadual do Oeste do Paraná, Cascavel, Brasil) - msoyamada@gmail.com ou marcio.oyamada@unioeste.br

Clodis Boscarioli, 0000-0002-7110-2026 (Universidade Estadual do Oeste do Paraná, Cascavel, Brasil) - boscarioli@gmail.com ou clodis.boscarioli@unioeste.br

A System for Information Extraction on Corporate Mobile Invoices

Information Extraction (EI) is a collection of methods and techniques that aim to extract relevant information from semi structured or unstructured sources. This paper proposes a supervised application for EI using the machine learning approach in Corporate Mobile Invoices. The system composed of modules for problem identification, processing and pattern extraction and post-processing was evaluated experimentally to measure its accuracy. The experiments showed results that may support the development systems capable of extracting information from mobile phones bills in general.

Keywords: Information Extraction, Machine Learning, Hierarchical Organization.

Um Sistema para Extração de Informação em Faturas de Telefonia Móvel Corporativa

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. Este artigo traz a proposta de uma aplicação supervisionada de EI utilizando a abordagem de aprendizado de máquina em Faturas de Telefonia Móvel Corporativa. O sistema, composto por módulos de identificação do problema, processamento, extração de padrões e pós-processamento, foi avaliado experimentalmente visando mensurar sua precisão. Os experimentos mostraram resultados que podem subsidiar desenvolvimento de sistemas capazes de extrair informações de faturas de telefonia móvel em geral.

Palavras-chave: Extração de Informação, Aprendizado de Máquina, Organização Hierárquica

1. INTRODUÇÃO

A complexidade e a quantidade de dados armazenados ou transmitidos no mundo têm aumentado a taxas elevadas nos últimos anos. O surgimento dessa enorme quantidade de dados constitui o que se chama de *Big Data*, sendo considerado o quarto paradigma da ciência, (Hitzler & Janowicz, 2013). No entanto, o grande volume aliado à falta de padrões torna inviável qualquer tipo de acompanhamento sistemático desses dados. Neste contexto, é cada vez maior a necessidade de métodos para a extração de informações neste tipo de dados.

Uma das soluções para este tipo de problema se dá por meio da aplicação de técnicas que permitem a estruturação da informação, tornando-a mais elucidativa para o usuário. Para isso, é necessário a construção de ferramentas capazes de processar esse grande volume de dados e que permita uma melhor análise de suas informações. Neste sentido, (Fraga do Amaral e Silva, 2004) sugerem uso de uma técnica de Mineração de Texto denominada Extração de Informação (EI), que tem como objetivo transformar documentos textuais não estruturados ou semiestruturados, em banco de dados estruturados.

A EI é um passo fundamental em vários problemas de análise de dados. Seus benefícios podem se estender a qualquer domínio que utilize algum tipo de texto (Loh, 2001), como é o caso da área de Gestão de Custos de Telecomunicações, de interesse neste trabalho. De acordo com (Cavalcante, 2009), o profissional desta área tem o objetivo de compreender, gerenciar e otimizar a utilização de serviços de telecomunicações dentro de um ambiente corporativo. Dada a complexidade nos processos e comercialização destes serviços, realizar este trabalho manualmente pode ser bastante custoso, sobretudo para grandes empresas. Para tanto, faz-se necessária uma interpretação automatizada das faturas telefônicas fornecidas pelas operadoras prestadoras de serviços.

As principais contribuições da EI neste tipo de documento estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de texto e uma melhor compreensão do conteúdo disponível nestes documentos.

Os sistemas de Extração de Informações têm o potencial de auxiliar os seres humanos na tarefa de extração, no entanto, a maioria dos sistemas de EI não foram projetados para funcionar com o formato PDF (*Portable Document Format*), uma fonte de extração importante e comum em várias áreas. Em um documento PDF, o conteúdo narrativo muitas vezes é misturado com metadados ou texto semiestruturados, o que adiciona desafios ao algoritmo de Mineração de Texto.

Com o objetivo de extrair informações relevantes de forma a melhorar as decisões estratégicas na área de Gestão de Custos de Telecomunicação, e automatizar o processo de estruturação das informações presentes em documentos textuais, este trabalho apresenta um sistema baseado em Extração de Informações em Faturas de Telefonia Móvel Corporativas em formato PDF.

Este artigo segue assim organizado: Uma visão geral de conceitos da Extração de Informações, além de processos, tarefas e técnicas que serão aplicadas neste trabalho são apresentados na Seção 2. A Seção 3 apresenta um estudo sobre algumas pesquisas desenvolvidas na área. A Seção 4 descreve a metodologia abordando as etapas da pesquisa como também a proposta de desenvolvimento do sistema de EI. A Seção 5 traz resultados

da avaliação experimental realizada e, por fim, na Seção 6 seguem as considerações finais e perspectivas da pesquisa.

2. MINERAÇÃO DE TEXTO E A EXTRAÇÃO DE INFORMAÇÃO

É possível encontrar na literatura diversas definições para o termo Mineração de Texto (MT). Os pesquisadores (Rezende, 2003) e (Feldman & Sanger, 2007) consideram a MT como sendo o processo de obtenção significativo de informações de texto no qual um usuário interage com uma coleção de documentos, usando um conjunto de ferramentas para análise.

A Mineração de Texto aplica as mesmas funções analíticas de Mineração de Dados, envolvendo 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 (Feldman & Sanger, 2007).

O processo de Mineração de Texto, segundo (Rezende, 2003), pode ser dividido em cinco etapas, como pode ser observado na Figura 1.

Figura 1: Processo de mineração de textos
Fonte: (Rezende, 2003)

A identificação do problema, consiste em determinar o escopo do problema, onde dado objetivo da aplicação é selecionada ou criada uma base de documentos, onde é definida o que se espera de sua análise. Esta etapa é fundamental para o decorrer do processo, pois o conhecimento adquirido servirá como base para etapas seguintes do processo.

O pré-processamento é uma etapa indispensável no processo de MT. Visa estruturar os textos de maneira a torná-los processáveis por algoritmos de aprendizado de máquina, padronizando ou convertendo textos em outros formatos ou eliminando *stopwords*, que são palavras consideradas irrelevantes à representação da coleção.

Na extração de padrões é escolhida a técnica de MD que será empregada para que os objetivos sejam atingidos. O pós-processamento visa avaliar as informações que foram

descobertas na etapa anterior, e a Utilização do Conhecimento consiste em consolidar o conhecimento e utiliza-lo para apoiar algum processo de tomada de decisão.

2.1 Tarefas da Mineração de Textos

Diversas tarefas existentes podem ser utilizadas para realizar a MT, como a Recuperação de Informação, Classificação, Agrupamento, entre outras. Uma estratégia comumente utilizada é a Extração de Informação, que segundo (Zambenedetti, 2002) começa com uma coleção de documentos textuais, de onde fragmentos de textos importantes são isolados, e informação relevante extraída. O objetivo da pesquisa em EI é construir sistemas que analisem um texto e preencha um modelo predeterminado com informações sobre o tipo específico de evento.

A escolha da técnica a ser usada na construção de sistemas extratores de informação dependem da formatação do texto de entrada (Fraga do Amaral e Silva, 2004). Entre as abordagens presentes na literatura estão as baseadas em *Wrappers*, que são definidas por regra ou padrões de extração para identificar os dados de interesse a serem extraídos.

Os *Wrappers* exploram a regularidade apresentada por textos estruturados ou semiestruturados, como Faturas de Telefonia Móvel Corporativas, a fim de localizar os dados importantes, e em geral exportá-las como parte de uma estrutura de dados.

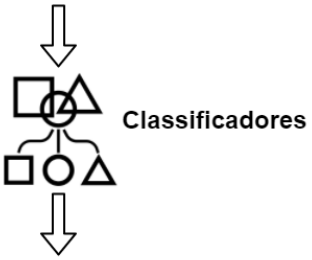
Por não existir uma arquitetura consensual para este tipo de abordagem, um *Wrapper* é construído de modo *ad-hoc*, existindo duas formas de implementação. A primeira, baseada em engenharia do conhecimento, na qual, uma pessoa familiarizada com o domínio elabora regras de extração de modo manual. A segunda é fundamentada em aprendizado de máquina que, diferente da primeira, não necessita de um indivíduo para codificar manualmente as regras, é preciso apenas que exista alguém com conhecimento suficiente do domínio, e da tarefa de extração, para etiquetar um *corpus* de textos de treinamento e teste. Tal etiquetagem consiste em determinar em cada texto as informações que deverão ser extraídas pelo sistema.

Conforme (Fraga do Amaral e Silva, 2004), um algoritmo de aprendizado de máquina pode ser executado, se um *corpus* de treinamento for criado de forma adequada, resultando em algum tipo de conhecimento que o sistema de EI poderá utilizar para realizar a extração das informações. Uma das principais técnicas utilizadas pelos *Wrappers* para EI é a Classificação.

2.2 Técnica de Classificação

Um classificador de texto é capaz de classificar uma entrada, representada por um vetor de características, em uma categoria ou classe predefinida. Contudo, dado um documento, não se deseja classificá-lo em um determinado tipo, e sim extrair alguns trechos que preenchem um formulário de saída desejado. Portanto, 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, como mostra a Figura 2. O fragmento que possuir maior grau de confiança é o escolhido para preencher o campo do formulário.

Realização de chamada Interurbana para o número (11)99999-9999, no período de 01/01/2017 às 14:15 com duração de 12:37s.



Número	Período	Horário	Duração
(11)99999-9999 100%	01/01/2017 100%	14:15 83%	12:37 75%

Figura 2: Exemplo de preenchimento de formulário com fragmento de texto

A partir dos fragmentos pode-se extrair uma série de características que formarão o vetor usado pelo classificador, as quais, segundo (Yang & Pedersen, 1997), podem ser extraídas dos conhecimentos do especialista no domínio ou de processos de seleção automática. Entretanto, segundo (Kushmerick & Thomas, 2003), uma das maiores limitações para se tratar de problemas de EI, é a geração de fragmentos relativamente significativos e corretos, 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.

Além disso, (Costa et al., 2007) demonstram que classificadores procuram suprir a necessidade de se identificar instâncias de forma mais rápida, usando características dos objetos como critério para classificação, possuindo dois modelos. O modelo discriminativo utiliza um padrão de classificação, fazendo uso dos atributos, para encontrar a classe do objeto. E o padrão probabilístico, que calcula a possibilidade de o objeto pertencer a uma dada classe.

Os classificadores Bayesianos são exemplos de classificadores probabilísticos (McCallum & Nigam, 1998). Segundo (Zhang, 2004), devido à sua simplicidade e alto poder preditivo os algoritmos Bayesianos, chamados de *naive Bayes*, comparados em seu trabalho, com os métodos de Árvores de Decisão e Redes Neurais Artificiais obtiveram resultados compatíveis. Tendo em vista sua fácil implementação e resultados satisfatórios, o algoritmo de *naive Bayes* foi utilizado nesse trabalho.

Em faturas de Telefonia Móvel Corporativas que possuem uma organização hierárquica conhecida previamente, é possível realizar uma abordagem específica, capaz de determinar a classe de cada um dos dados presentes nos documentos. A abordagem aplicada, e denominada Classificação por Nó Pai, consiste em treinar um classificador tradicional *naive Bayes* para cada classe pai da hierarquia, como mostra a Figura 3, onde cada classificador é representado por um retângulo pontilhado em torno dos nós.

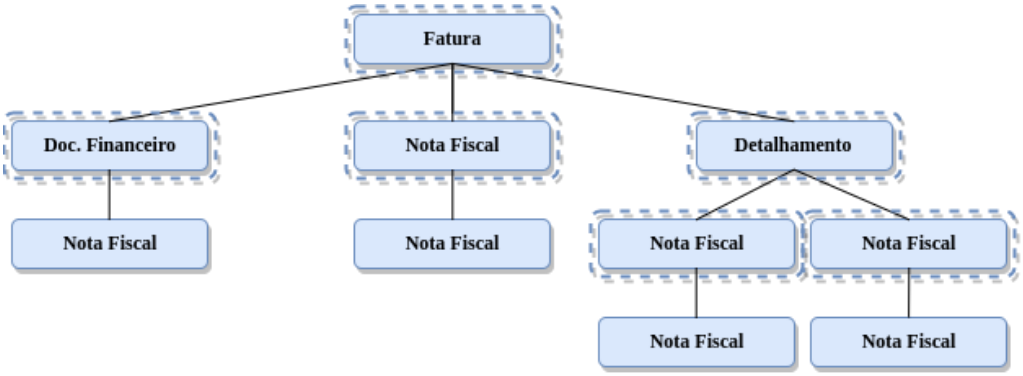


Figura 3: Abordagem de classificação por Nó Pai

Essa abordagem é fundamental na hipótese de que, quando utilizados diferentes algoritmos de classificação em cada nó da hierarquia, a acurácia preditiva é melhorada (Secker et al., 2007), visto que cada nó da árvore apresenta um comportamento referente a diferentes informações dispostas na fatura telefônica, ou seja, o nó "Nota Fiscal" apresentado na Figura 3, contém apenas informações pertencentes a seção de Notas Fiscais da fatura, em razão disso, quando este nó é utilizando individualmente, não gera conflito com os demais nós da arvore, elevando portanto a acurácia preditiva do algoritmo de classificação deste e dos demais nós.

3. TRABALHOS CORRELATOS

No trabalho realizado em (Kumar et al., 2018), é proposto um novo método de Extração de Informações para identificação de varreduras de entidades espaciais e relacionamentos entre dois tipos de textos curtos. A pesquisa foi realizada utilizando especificamente textos curtos (SMS e *tweets*), escritos eletronicamente no idioma francês. Inicialmente, os autores propõem uma medida de similaridade, para melhorar a análise estatística da tarefa de reconhecimento de texto. Depois identificam novas relações espaciais, combinando abordagens de Processamento de Linguagem Natural (PLN). Seus resultados mostram que houve uma melhora na qualidade da extração de novas relações espaciais em textos do idioma francês. Porém, seu processo ainda possui limitações acerca da aplicação em textos de outros idiomas.

A Extração de Informações em documentos de formato PDF, foram pesquisados em (Bui & Del Fiol & Jonnalagadda, 2016). Nesse trabalho, os autores falam sobre a complexidade de se extrair informações de documentos em formato PDF, que muitas vezes apresentam informações semiestruturadas, misturadas com metadados e informações importantes. Seu objetivo é categorizar o PDF em formato textual para uso estratégico por sistemas de EI. Sua metodologia é baseada em uma ferramenta de código aberto para realizar a extração de texto bruto de um documento. Além disso, foi desenvolvido um algoritmo de classificação que segue uma estrutura de passagem múltipla para classificar automaticamente fragmentos de texto, em 5 categorias. Seus testes, utilizando um algoritmo de peneira de passagem múltipla alcançou uma precisão de 92,6%, comparado com algoritmos de aprendizagem de máquinas que utilizam regressão logística.

No contexto de classificação, trabalhos que possuem classes organizadas hierarquicamente podem ser encontrados em diversas áreas de aplicação, como é o caso da Bioinformática, em (Nievola et al., 2014) que visa classificar proteínas em classes funcionais, que se encontram organizadas hierarquicamente, em aplicações de

reconhecimento de imagens, objetos também podem ser classificados por suas relações de dependência, como nos trabalhos de (Vens et al., 2008) e (Kumar et al., 2018). Na área de classificação de documentos, textos podem ser caracterizados por sua hierarquia de assuntos ou tópicos, como em (Sun & Lim, 2001) e (Silla & Freitas, 2011).

Notou-se também a grande utilização de Extração de Informação no contexto de obtenção de informações em documentos de artigos científicos, como é o caso de trabalhos de (Fraga do Amaral e Silva, 2004) e (Álvarez, 2007), no entanto, não se encontrou trabalhos relacionados à área de telecomunicações, mais precisamente em faturas de Telefonia Móvel Corporativas, propósito deste trabalho.

4. METODOLOGIA

Conhecida a abordagem utilizada para classificar os itens de uma fatura, a definição do sistema proposto, como pode ser visto na Figura 4, é dividido sequencialmente em dois processos, a Identificação do Problema e a Extração de Informação. As etapas serão descritas ao longo desta seção.

O primeiro processo tem como principal objetivo, a coleta do *corpus* de documentos (1.1), que foi realizada mediante a disponibilização por um especialista no domínio do problema, que trabalha com esses tipos de documentos, e que necessita de um sistema para auxiliá-lo.

Ressalta-se que, o uso de faturas corporativas é justificado por apresentarem uma grande quantidade de conteúdo e os relacionar de forma hierárquica, sendo possível distinguir conceitos, de maior e menor relevância, se comparadas às faturas de uso convencionais e não corporativas. Elas também possuem uma organização complexa, com regras distintas, que muitas vezes acabam sendo modificadas pela não padronização na organização dos dados por parte das operadoras de telefonia.

Para (Cavalcante, 2009), a fatura telefônica trata-se de um documento complexo pelas seguintes razões: São excessivamente técnicas, possuem um padrão próprio, estão enquadradas em uma legislação específica e refletem a complexa estrutura da operação.

Estas razões justificam a escolha de apenas uma operadora para a construção do modelo proposto. 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.

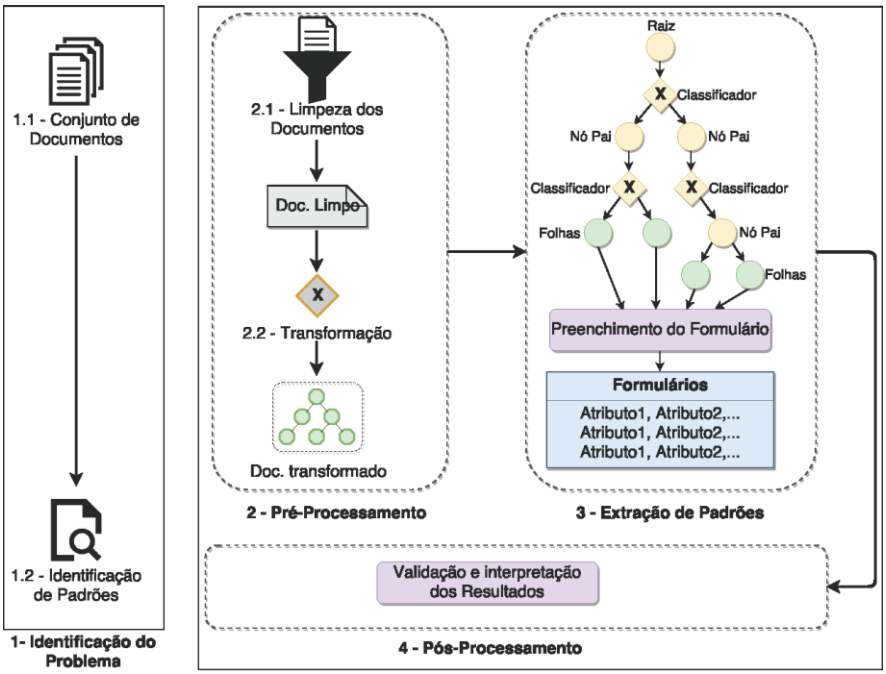
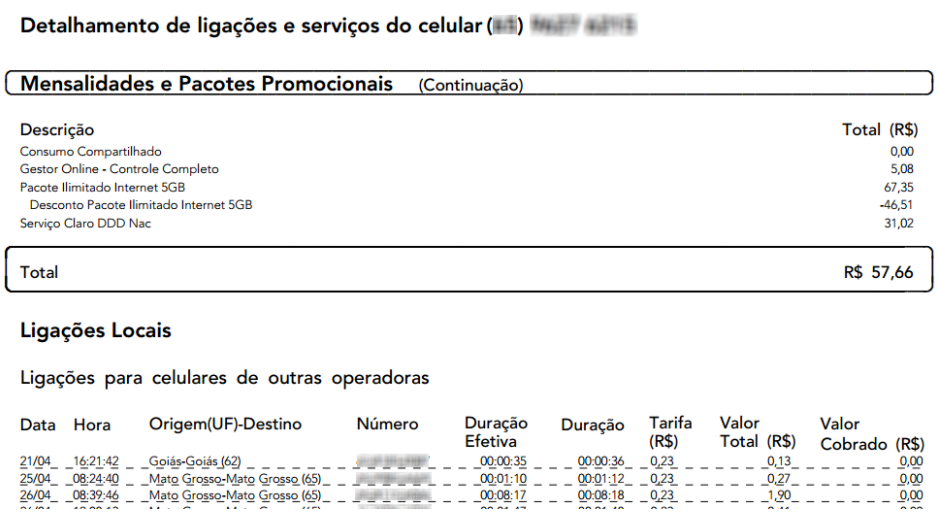


Figura 4: Modelo proposto para elaboração da Extração de Informações

A partir de uma fatura telefônica móvel corporativa, podem ser extraídas diversas informações, como valores monetários, horários de utilização, números de origem e destino, o tipo de ligação, entre outros dados, como pode ser visto na Figura 5.



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 5: Exemplo de 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, associados aos serviços contratados, ou seja, dentro dessa categoria, pode-se citar as contas de consumo em gral, que têm uma parcela mínima a ser paga todo mês, e gastos variáveis, associados a consumos não previstos na contratação e que são tarifados de forma avulsa, gerando um gasto acima do contratado.

Com o objetivo de obter mais informações 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, já o detalhamento de cada item traz informações

como horário em que item foi consumido, o número que o consumiu, origem e destino do consumo, entre outros.

A extração das informações contidas nestes detalhamentos possui como principal motivação a criação de uma base de dados (formulário de saída), como mostra a Tabela 1 capaz de fornecer uma ferramenta útil para serviços de auditoria e consultoria na área de Gestão de Custos em Telecomunicação, trazendo mais praticidade e qualidade uma vez que o profissional da área precisa consultar e analisar tais informações de forma manual.

Tabela 1: Exemplo de formulário de saída

Campo	Descrição
Tipo	Tipo do detalhamento apresentado
Descrição	Descrição do item extraído da própria linha
Modalidade	Modalidade utilizada (Voz, Dados, SMS)
Tarifa	Valor da tarifa praticada do consumo
Valor Cobrado	Valor cobrado pelo serviço

É com base nessas informações, que ainda no primeiro processo, é realizada a identificação e exploração de padrões interessantes nos documentos (1.2). Esses padrões são utilizados como base para a elaboração das regras, na construção dos classificadores nas etapas seguintes, e na escolha da técnica mais adequada à tarefa de Extração de Informação.

O texto dos documentos é do tipo semiestruturado, ou seja, a forma como os dados estão organizados não estão de acordo com uma estrutura formal se comparados com bancos de dados relacionais. Eles ainda possuem características de: Campos Ausentes, pois em geral não possuem informações para preencher todos os campos do formulário de saída, Estilo Telegráfico, em que muitas palavras aparecem abreviadas, Ausência de Delimitadores Precisos, não existindo delimitadores de onde começa ou termina cada campo a ser extraído e a Hierarquia de Tópicos, visto que faturas telefônicas apresentam em seu conteúdo uma hierarquia de itens que devem ser levados em consideração para que se consiga obter corretamente as informações desejadas.

Todas essas características, em especial a característica de hierarquia motivou a elaboração do modelo proposto, que ocasionou na escolha da técnica de Classificação Hierárquica.

Após realizada a primeira etapa do processo, o modelo proposto segue com a Extração de Informação. Esse processo engloba as etapas de Pré-processamento (2), onde se realiza a limpeza (2.1) e transformação (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.

4.1 Extração de Informação e Etiquetagem

As técnicas escolhidas para a realização do Pré-processamento, e da Extração de Padrões, utilizam respectivamente a classificação plana e a classificação hierárquica. Ambas as classificações, são processos supervisionados e necessitam de um *corpus* etiquetado de faturas para treinamento de seus classificadores, a fim de alcançar o objetivo definido na tarefa a ser executada.

Ambos os processos de classificação requerem que os documentos sejam divididos em várias instâncias de texto (linhas da fatura), que serão submetidos primeiramente à etapa de Pré-processamento, para avaliar se as mesmas são relevantes ou irrelevantes para

o sistema, e posteriormente ao processo de Extração de Padrões, a fim de extrair da instância as informações que deverão preencher o formulário de saída.

O *corpus* etiquetado de fatura utilizado, 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 Padrões.

A etiquetagem foi realizada utilizando engenharia do conhecimento e tomou como base o estudo realizado na primeira etapa do processo. As ferramentas e tecnologias utilizas para a construção do sistema, bem como para realizar a etiquetagem do corpus, foram o SGBD PostgreSQL (PostgreSQL, 2005), utilizado para armazenar as informações no banco de dados, a linguagem de programação Java, bem como a biblioteca *iText* (iText Working Group, 2001), usada para construção, leitura e manipulação de documentos em formato PDF.

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 Padrões. 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 escolhidas com a ajuda do especialista no domínio, e que proporcionam a realização do pré-processamento são: Nível-1, Nível-2, Nível-3, Nivel-4, Folha. Essas classes têm como objetivo a representação dos graus mais externos, aos mais internos da fatura, respectivamente. Já a classe Lixo, representa instâncias que possuem características que as diferenciam das linhas importantes da fatura.

A distinção entre os itens de Níveis 1, 2 e 3, foi realizada 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 bastante semelhante. Um exemplo desta complexidade pode ser visualizado na Tabela 2, onde as instâncias em cinza representam Lixos, e as instâncias em branco representam dados relevantes a análise. É possível, perceber que as linhas possuem grande similaridade, e para diferenciá-las foi necessário o uso de características mais detalhadas.

As linhas que possuem por exemplo o caractere "$" foram consideradas pelo especialista, como não relevantes para a EI, bem como instâncias que apresentam dois espaços consecutivos ou que apresentam espaços no final de sua cadeia de caracteres.

Tabela 2: Amostra de similaridade entre classes Lixo e Folha

1	Ligações para celulares Claro #,# #,# #,# #,#
2	Ligações para celulares Claro #,# #,# #,# - #,#
3	Desconto Pacote Ilimitado Internet #GB - - - - R$ -#,#
4	Desconto Pacote Ilimitado Internet #GB -#,# -#,# - -#,# - -#,#
5	Pacote Ilimitado Internet #GB -#,#
6	Pacote Ilimitado Internet #GB #,#
7	Assinatura Plano Sob Medida R$ #,# R$ #,# R$ #,# R$ #,#

8	Assinatura Plano Sob Medida #,# #,# #,# - #,#

Após a compreensão das instâncias, as que obtiveram identificação de Lixo, foram descartadas das próximas etapas de EI. Na etapa de definição das classes de informação, foi acrescentada uma etiqueta que corresponde a 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 atentar que essa etiqueta não identifica diretamente o campo a ser preenchido, e sim informações que podem ser usadas como base para preencher diversos campos. Exemplos de etiquetas utilizadas podem ser visualizadas na Tabela 3.

Tabela 3: Exemplo de etiquetas usadas na classe de informação

Etiqueta	Descrição
A cobrar-recebida	Identifica uma ligação que foi recebida de forma a cobrar.
Consumo	Consumos de forma geral que foram realizados na fatura.

Deve-se lembrar que cada linha pode ser encontrada em várias seções diferentes da fatura, ou seja, pode possuir mais que um pai, e pode possuir várias classes associadas à sua classe original. Um exemplo deste tipo de etiquetagem pode ser visualizado na Figura 6.

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

Como pode ser visto na Figura 6, alguns itens possuem mais de uma classe original associada, essas classes são separadas por vírgulas para melhor identificação nas próximas etapas. Deve-se atentar 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 pode ser visualizado na Figura 7.

	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 7: Exemplo de etiquetagem utilizando a classificação de forma hierárquica em linhas do tipo folha

Na Figura 7 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 adequadamente.

4.2 Pré-Processamento

O corpus utilizado pode conter faturas que foram geradas pelas operadoras, 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. Logo, nesta fase, os dados normalmente dispostos em formato inadequado, são preparados para que possam ser utilizados pelos algoritmos de extração de padrões. Diversas tarefas são executadas nesta etapa, como a transformação e redução de dados e limpeza.

A Transformação, tem como finalidade, superar quaisquer limitações presentes nos algoritmos de extração de padrões, por esse motivo foi necessário modificar a representação dos dados. Como a correção de erros provenientes da leitura das faturas utilizando a biblioteca *iText*.

Já na Redução, em função de limitações de memória e tempo de processamento, foi necessário aplicar métodos para redução de dados, como a criação de uma máscara, para substituir os itens em formato numérico por um caractere "#". Exemplo: (35,14 em #,#). Esse processo, auxilia na diminuição da quantidade de linhas iguais apresentadas na fatura, e que são armazenadas no banco de dados, como base de conhecimento.

Na Limpeza, os dados coletados podem apresentar diversos problemas, entre eles linhas com valores desconhecidos, incorretas ou inválidas, sendo importante a realização de uma limpeza nestes dados. Para realizar esta etapa, e identificar as linhas relevantes no processo de EI, utilizou-se um classificador *bayesiano* linear. Esse processo requisitou que os documentos fossem divididos em várias instâncias de texto (linhas das faturas), as quais primeiramente foram extraídas características destas instâncias, a fim de realizar os cálculos para determinar se as mesmas eram relevantes ou irrelevantes para o sistema.

A extração das características utilizadas neste processo foi baseada 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.

Tabela 4: Identificação de características para o classificador de limpeza

Característica	Descrição da característica
Títulos	Contém padrão de título (Nota Fiscal, Documento Financeiro)
Tem-Cifrão	Possui ao menos um símbolo de cifrão
Alfa-Numérico	Contém tanto letras quanto números
Tem-Sequência-EouF	Mais que dois “E” ou “F” consecutivos
Tem-Padrão-Lixo	Possui padrão de lixo como sequência de letras que forma o código de barras

Os resultados obtidos em relação ao algoritmo de classificação para a etapa de limpeza, estão especificados na Seção 5.

É importante observar que diante do resultado obtido, 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.

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.

5. EXPERIMENTOS E RESULTADOS

Os dados utilizados nos experimentos deste trabalho, foram disponibilizados pela empresa Orb it Sistemas, situada na cidade de Cascavel-PR, sendo um total de 1200 faturas telefônicas em formato PDF.

Inicialmente para treinamento do algoritmo de pré-processamento foram utilizados 1100 documentos, do total do *corpus* disponibilizado. 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 precisão entre os itens classificados. Este resultado foi gerado por meio da comparação com uma base já classificada manualmente pelo engenheiro do conhecimento.

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.

Os testes referentes a etapa de extração de padrões foram realizados inicialmente utilizando um composto de 800 faturas para treinamento e 5 faturas para testes, de modo que todos os documentos foram selecionados de forma totalmente aleatória do *corpus* de documentos, assim como o teste de pré-processamento.

O critério de avaliação do sistema de extração foi baseado em duas variáveis: *N-relevantes*, que expressa o número de informações relevantes identificadas previamente com a ajuda do especialista do domínio que devem ser extraídas da fatura, e *N-corretos*, que 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 à relação percentual entre a quantidade de informações que devem ser extraídas dos documentos.

No primeiro experimento com objetivo de avaliar a precisão final do sistema de EI utilizando a classificação hierárquica, fez-se uso de uma quantidade significativamente grande de documentos para treinamento.

Já o segundo experimento tem o intuito de demonstrar a precisão do sistema utilizando um conjunto de treinamento menor que o apresentado no primeiro. Essa segunda etapa utiliza uma variação na quantidade de documentos usados na etapa de treinamento, alternando a quantidade de faturas em 800, 400, 200 e 100, respectivamente aplicadas a cada um dos 5 documentos utilizados nos testes, como mostra a Tabela 5. Atenta-se que, a escolha desses documentos, assim como no primeiro experimento, foi realizada por meio de uma função de aleatoriedade aplicada ao sistema.

A redução na quantidade de documentos usados para treinamento, tem como objetivo, mostrar que mesmo com uma diminuição significativa da quantidade de instâncias de treino, os classificadores hierárquicos mostram-se promissores para a tarefa de extração de informação em faturas de telefonia móvel corporativas.

Tabela 5: Precisão média do sistema utilizando 5 faturas como teste

Documentos de treinamento	Instâncias de treino	Total de Instâncias de Teste	Precisão
800	13.314.227	19.040	100%
400	8.284.116	19.040	100%
200	3.456.758	19.040	≅100%
100	1.358.057	19.040	≅97%

O Experimento 1 mostrou um desempenho significativo do sistema utilizando a mesma quantidade de documentos para treinamento em ambos os cinco casos de teste, totalizando um resultado de 100% na extração das informações.

O Experimento 2 mostrou a precisão obtida pelo sistema com diferentes tamanhos do conjunto de treinamento, observando que a taxa de acerto do sistema se manteve alta utilizando uma quantidade relativamente menor de instâncias de treino.

6. CONCLUSÃO

As organizações empresariais buscam alternativas para melhorar seus processos de negócio. Uma delas é a aplicação de técnicas computacionais a fim de que colaborem de forma eficiente no processo de otimização de tempo e resposta para seus clientes. Neste contexto, torna-se necessário a utilização de mecanismos computacionais que permitam a interação com estes dados de forma inteligente e rápida.

A principal contribuição deste trabalho foi a utilização de uma abordagem de classificação hierárquica para a extração de informações em faturas de telefonia móvel corporativa por meio de um *wrapper* que utilizou um conjunto de classificadores probabilísticos.

Foram realizados diversos experimentos que mostram que, no domínio de documentos de faturas telefônicas, esta abordagem consegue extrair informações especificadas pelo usuário com precisão. Os resultados obtidos pela execução do sistema, mostram-se promissores em relação a técnica abordada. Além disso, um sistema para EI em faturas telefônicas foi desenvolvido, validado por um especialista no domínio, e que pode ser utilizado na prática em empresas que realizam a gestão dos serviços de telecomunicações, a fim de auxiliar no processo de tomada de decisão.

Estudos de técnicas de extração de informação que fazem uso de classificadores, foram realizados, 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.

A complexidade de se extrair informações para a elaboração das características utilizadas pelos classificadores a fim de determinar os itens relevantes e não relevantes dos documentos, tornou-se uma dificuldade, devido ao fato de uma fatura telefônica possuir uma estrutura não regular, e linhas bastante similares em seus tópicos internos.

Algumas melhorias além de aprimorar os resultados parciais do processo de extração atual, podem facilitar o uso e minimizar o tempo da atividade, como o aprimoramento da etapa de limpeza utilizando Modelos de Markov Ocultos (*Hidden Markov Models* - HMM), e a aplicação de técnicas como o Cálculo de Funções de Distância, para verificar o grau de similaridade em textos.

REFERÊNCIAS

Álvarez, A. C. (2007). *Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem* (Doctoral dissertation, Universidade de São Paulo).

Bui, D. D. A., Del Fiol, G., & Jonnalagadda, S. (2016). PDF text classification to leverage information extraction from publication reports. *Journal of biomedical informatics*, *61*, 141-148.

Cavalcante, J. R. R. (2009). *Gestão de custos em telecom*. Editora E-papers.

Costa, E., Lorena, A., Carvalho, A. C. P. L. F., & Freitas, A. (2007). A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods for machine Learning II: papers from the AAAI-2007 Workshop* (pp. 1-6).

Fraga do Amaral e Silva, E. (2004). Um sistema de extração de informação em referências bibliográficas baseado em aprendizagem e máquina. Dissertação de mestrado. Universidade Federal de Pernambuco, Recife, PE, Brasil.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, *4*(3), 233-235.

iText Working Group, Geisler, C., Bazerman, C., Doheny-Farina, S., Gurak, L., Haas, C., et al. (2001). IText: Future directions for research on the relationship between information technology and writing. *Journal of Business and Technical Communication*, *15*(3), 269-308.

Kumar, V., Pujari, A. K., Padmanabhan, V., Sahu, S. K., & Kagita, V. R. (2018). Multi-label classification using hierarchical embedding. *Expert Systems with Applications*, *91*, 263-269.

Kushmerick, N., & Thomas, B. (2003). Adaptive information extraction: Core technologies for information agents. In *Intelligent information agents* (pp. 79-103). Springer, Berlin, Heidelberg.

Loh, S. (2001). *Abordagem baseada em conceitos para descoberta de conhecimento em textos. 110 f* (Doctoral dissertation, Tese (Doutorado em Ciência da Computação)–Instituto de Informática, UFRGS, Porto Alegre).

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).

Nievola, J.C.;Paraiso, E.Malucelli, A.; Kaestner, C.; Enembreck, F.; Steiner, M. T. A. (2014). *Construção de Classificadores Hierárquicos Multirrótulo usando Evolução Diferencial*. Tese (Doutorado em Programa de Pós-Graduação em Informática) - Pontifícia Universidade Católica do Paraná.

PostgreSQL, D. (2005). Disponível em: http://www.postgresql.org. *Acesso em*, *15 de fevereiro de 2018*.

Rezende, S. O. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda.

Secker, A. D., Davies, M. N., Freitas, A. A., Timmis, J., Mendao, M., & Flower, D. R. (2007). An experimental comparison of classification algorithms for hierarchical prediction of protein function. *Expert Update (Magazine of the British Computer Society's Specialist Group on AI)*, *9*(3), 17-22.

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*(1-2), 31-72.

Sun, A., & Lim, E. P. (2001). Hierarchical text classification and evaluation. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on* (pp. 521-528). IEEE.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., & Blockeel, H. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning*, *73*(2), 185.

Yang, Y., & Pedersen, J. O. (1997, July). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412-420).

Zambenedetti, C. (2002). *Extração de informação sobre bases de dados textuais*. Dissertação (mestrado em Programa de Pós-Graduação em Computação) - Universidade Federal do Rio Grande do Sul.

Zhang, H. (2004). The optimality of naive Bayes. *AA*, *1*(2), 3.