UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS QUIXADÁ
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

PEDRO JOSÉ RODRIGUES DE SOUSA

UMA PROPOSTA PARA ANÁLISE DE DADOS ABERTOS DO PROCON UTILIZANDO DATA MART

QUIXADÁ
2014

PEDRO JOSÉ RODRIGUES DE SOUSA

UMA PROPOSTA PARA ANÁLISE DE DADOS ABERTOS DO PROCON UTILIZANDO DATA MART

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Sistemas de Informaçãoda Universidade Federal do Ceará como requisito parcial para obtenção do grau deBacharel.

Área de concentração: computação

Orientadora Prof[a]. Ticiana Linhares Coelho da Silva

QUIXADÁ
2014

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca do Campus de Quixadá

S696p Sousa, Pedro José Rodrigues de
Uma proposta para análise de dados abertos do PROCON utilizando Data Mart / Pedro José Rodrigues de Sousa. – 2014.
41 f. : il. color., enc. ; 30 cm.

Monografia (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2014.
Orientação: Prof. Me. Ticiana Linhares Coelho da Silva
Área de concentração: Computação

1. Mineração de dados (computação) 2. KDD (Recuperação da informação) 3. Sistemas de computação - Tecnologia da Informação I. Título.

CDD 005.3

PEDRO JOSÉ RODRIGUES DE SOUSA

UMA PROPOSTA PARA ANÁLISE DE DADOS ABERTOS DO PROCON UTILIZANDO DATA MART

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Sistemas de Informação da Universidade Federal do Ceará como requisito parcial para obtenção do grau de Bacharel.

Área de concentração: Computação

Aprovado em: _____ / junho / 2014.

BANCA EXAMINADORA

_____
Profa. MSc. Ticiana Linhares Coelho da Silva
(Orientadora)
Universidade Federal do Ceará-UFC

_____
Prof. MSc Regis Pires Magalhães
Universidade Federal do Ceará-UFC

_____
Prof. MSc. Camilo Camilo Almendra
Universidade Federal do Ceará-UFC

Dedico este trabalho as pessoas que sempre me fizeram acreditar na realização do meu sonho, e me apoiaram para que eu pudesse realizá-lo, meus pais, esposa, filhos, amigos, professores, que estiveram presentes, durante todos os momentos difíceis e que compartilham comigo desta alegria.

AGRADECIMENTOS

Registro aqui meus agradecimentos, a todos os que participaram direta ou indiretamente da caminhada que trilhei, para que eu realizasse este trabalho, auxiliando-me e dando-me forças nos momentos em que mais precisei.

Agradeço a Deus, por estar comigo em todos os momentos e iluminando-me, sendo meu refúgio e fortaleza nos momentos mais difíceis. A ele, minha eterna gratidão.

Agradeço, especialmente, à minha família, por me apoiar para que pudesse concretizar esse trabalho: minha mãe e meu pai, que foram incansáveis; e, em especial, minha esposa, que esteve ao meu lado, tentando entender-me nos momentos de ausência, dando-me apoio e carinho.

À professora mestre Ticiana Linhares Coelho da Silva, minha "orientadora, companheira e amiga", que me possibilitou "aprendizagens únicas" que levarei comigo por uma vida inteira, por meio do grande incentivo e orientação que me foram concedidos durante essa jornada.

Aos amigos que fiz dentro da Universidade Federal do Ceará, e a todos os que fazem parte desta comunidade.

“A informação muda a natureza da competição, pois você não pode mais beneficiar-se da ignorância do consumidor.” (William Bill Gates)

RESUMO

Atualmente, é gerada uma grande quantidade de dados por sistemas de informação. Uma parte destes dados é disponibilizada ao público em geral, através de iniciativas de alguns órgãos. No entanto, a maioria desses dados não são disponibilizados de maneira legível, apresentam incoerências e inconsistências. Assim, é necessário propor soluções que manipulem esses dados e os tornem compreensíveis para que a informação seja transmitida ao público. Este trabalho utilizou os dados abertos da Fundação de Proteção e Defesa do Consumidor (PROCON), disponível ao público no portal de Dados Abertos do Governo Federal. Para transformar esses dados em informações úteis, foi realizado o processo de descoberta de conhecimento em bases de dados(ou KDD – *Knowledge Discover in Databases*, em inglês) e foi proposto um modelo ROLAP (*Relational On Line Analytical Processing*) para construção de um Data Mart em que as análises por meio de consultas OLAP(*On-Line Analytical Processing*) possam ser realizadas. Os resultados dessas consultas permitiram a geração de gráficos que facilitam o entendimento das informações contidas. O povoamento do *Data Mart* foi realizado utilizando o *Pentaho Data Integration*, que consiste em ferramenta de ETL (*Extract, Transform, Load*). As consultas OLAP realizadas e os gráficos gerados, foram construídos utilizando-se O *Pentaho Business Analytics*, uma ferramenta bastante utilizada no mercado de Tecnologia da Informação e que garante a corretude das análises realizadas. Ainda sobre a análise dos dados, o autor tentou criar um sistema de recomendação com base no perfil dos usuários do PROCON. Porém, não foi possível estabelecer o perfil dos usuários utilizando tais dados abertos. Como trabalhos futuros, planeja-se a coleta de novos dados para análises temporais, e verificação da possibilidade de construção de um sistema de recomendação com o cruzamento de outras bases de dados abertos referentes ao PROCON. Por fim, o presente trabalho foi apresentado como Minicurso no Workshop de Tecnologia da Informação do Sertão Central (WTISC 2014) realizado pela Universidade Federal do Ceará, Campus Quixadá.

Palavras chave: Legibilidade. Análise. Dados Abertos.

ABSTRACT

Nowadays, it has been generated a large amount of data. There are many government systems that provide these data public. However, most of these data available are not legible, and they are incoherent and inconsistent. Thus, it is necessary to propose solutions that manipulate these data, make them understandable as information, and provide them to the public community. This study used data from the Fundação de Proteção e Defesa do Consumidor (PROCON), available to the public on Portal de Dados Abertos do Governo Federal. To transform this data into useful information, we performed all the process of Knowledge Discovery in Databases (KDD) and we proposed a Data Mart using ROLAP (Relational Online Analytical Processing) as a model for the construction. The tests through OLAP (On-Line Analytical Processing) queries were processed on the Data Mart proposed. The results of these queries enabled the generation of graphs that facilitate the understanding of the information. Still on the data analysis, the authors attempted to create a recommendation system based on user profiles of PROCON. However, it has not been possible to establish the profile of users using such open data. As future work, we plan to collect new data for more time analysis, and verify the possibility of create a recommendation system with the integration of other databases related to PROCON. Finally, this work was presented as short course on the Workshop de Tecnologia da Informação do Sertão Central (WTISC 2014) that occurred on Federal University of Ceará, Campus Quixadá.

Keywords: Legibility. Analysis. Open Data.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Nas duas últimas décadas, o aumento contínuo do poder computacional tem produzido um fluxo enorme de dados (JI et al., 2012). A cada dia, 2,5 quintilhões de bytes de dados são criados e 90% dos dados no mundo hoje foram produzidos nos últimos dois anos (WU, et al. 2013). Sistemas corporativos, serviços e sistemas Web, redes sociais, transações financeiras, *e-commerce* entre outros, produzem juntos um grande volume de dados, alcançando a escala de petabytes diários. Estes exemplos exigem armazenamento eficiente, além da necessidade de extrair conhecimento por meio de análises que auxiliem no processo de tomada de decisão, por exemplo.

Não são todos esses dados disponíveis na Web que estão em forma legível a qualquer usuário da Internet. Alguns dados somente podem ser visualizados, algumas vezes por uma parcela dos profissionais de TI, já que os mesmos todos possuem conhecimentos suficientes e se dedicam a tal atividade. Pois os mesmos já possuem conhecimento sobre o potencial de informação que pode ser extraído dessas bases de dados.

Um grande e importante serviço oferecido na Internet é a disponibilidade de dados sobre órgãos governamentais, a fim de proporcionar uma transparência sobre seus serviços ao público em geral. É importante ressaltar que nem todos os dados governamentais estão disponibilizados ao público, e também não são apenas dados governamentais que estão abertos ao público, existem uma variedade enorme de dados disponibilizados na web. Somente alguns órgãos ou entidades tomam a iniciativa de publicá-los. Por exemplo, os dados disponibilizados pelo TCM-CE (Tribunal de Contas dos Municípios do Ceará) que visam informar aos cidadãos de que forma é gasto o dinheiro público, para que os mesmos possam ser atuantes na fiscalização do trabalho exercido por seus gestores.

Os dados governamentais são um dos tipos de dados abertos disponíveis na web. Dados abertos são dados legíveis por máquina e estão disponíveis livremente de modo que qualquer pessoa possa usá-los, reutilizá-los e redistribuí-los, tendo a exigência apenas de creditar a sua autoria e compartilhar pela mesma licença (OKFN, 2004). No tocante a legibilidade destes dados, é necessária a iniciativa de profissionais de TI para realizar trabalhos específicos, a fim de tornar os mesmos compreensíveis a todos os usuários da internet utilizando softwares que possam transformar esses dados em informações relevantes. Uma forma comumente utilizada para obter essas informações é consultar em grandes bases de dados por meio de aplicações OLAP.

As aplicações OLAP (*On-Line Analytical Processing*) normalmente acessam grandes bases de dados, realizando consultas intensivas. As atualizações podem ocorrer, mas em horários predefinidos específicos. Além disso, as consultas OLAP têm uma natureza *ad-hoc* e são utilizadas para análise e extração de conhecimento (LIMA, 2004).

Outra possibilidade de extração de conhecimento eficiente dos dados pode ser obtida a partir de técnicas de mineração de dados. Tais técnicas podem ser utilizadas para analisar e entender os dados a serem manipulados. A análise é baseada em modelos capazes de sumarizar dados, extrair novos conhecimentos ou realizar predições. Estes modelos podem ser utilizados para construir um software que possibilite, por exemplo, identificar o perfil de clientes para conceder empréstimos bancários, aplicações de recomendação de busca de amigos em redes sociais, que envolvem grafos com milhões de nós e arestas ou, ainda, sistemas de software que identifiquem possíveis ameaças terroristas (SILVA, 2013; RAJARAMAN, ULLMAN, 2012).

Este trabalho visa extrair informações importantes dos dados abertos do PROCON, disponibilizados pelo Governo Federal no portal Dados Abertos. O objetivo é propor um modelo de *Data Mart* e realizar análise destes dados, utilizando algumas técnicas de extração de conhecimento, além de permitir a visualização dessas informações em forma legível a qualquer cidadão, por meio de gráficos.

Durante o processo de realização deste trabalho, os dados foram coletados e avaliados quanto às possíveis inconsistências existentes nos mesmos, impureza e incompletude, inclusive o resultado da análise desses problemas podem ser enviados como feedback aos responsáveis pela sua publicação, para que possam melhorar a qualidade da publicação de futuras versões dos mesmos dados. Posteriormente, foram definidas as análises/consultas a serem realizadas sobre os dados. Duas metodologias de como analisar tais dados, OLAP e mineração de dados, foram investigadas. Além disso, foi proposta uma modelagem para a base de dados em questão, a fim de facilitar a realização de consultas, considerando que a base pode referir-se à vários anos de reclamações.As informações extraídas dos dados serão disponibilizadas neste trabalho por meio de gráficos, disponibilizadas ao público através deste trabalho acadêmico ou de trabalhos futuros, a fim de facilitar a compreensão das informações pela população em geral, das empresas citadas nas reclamações ou ainda órgãos públicos fiscalizadores.

As informações poderão ajudar as empresas a analisar o perfil de seus consumidores insatisfeitos, dando assim um importante *feedback* para que estas venham a

tomar melhores decisões sobre seus produtos, serviços, atendimento e traçar estratégias de melhoria para o seu mercado consumidor, atendendo de forma eficiente a seus clientes.

Atualmente, estão disponibilizadas duas aplicações sobre dados abertos do PROCON. Tais aplicações permitem que as pessoas interessadas possam ter acesso a algumas informações, como por exemplo, é permitido verificar qual a quantidade de reclamações feitas pela população a cada empresa. Estes aplicativos estão disponíveis no site de dados abertos do governo na guia (APLICATIVOS, 2013). No entanto, tais aplicações apresentam algumas restrições de informações e sobre o ano de pesquisa, restringindo-se a anos anteriores a 2012. Em específico, o aplicativo Reclamações PROCON aborda somente do ano de 2011. Já o aplicativo Reclamações BR que classifica empresas por reclamações, mostrando o pior índice de solução de atendimentos, manipula dados dos anos de 2009 a 2011. Porém, a base de dados não é a mesma da utilizada neste trabalho.

A solução proposta neste trabalho busca provê uma quantidade maior de informações relevantes ao usuário, de tal sorte que se torne possível realizar diversas análises, em qualquer ano em que forem disponibilizados os dados no portal de dados abertos, não se atendo apenas a um ano específico, como o aplicativo Reclamações PROCON. É possível visualizar quais serviços estão sendo oferecidos a população com baixa qualidade, que nesse caso será medido pela quantidade de reclamações, realizando análises por sexo, idade, região, estado, cidade, mês, ano, trimestre, produtos, entre outros. Essas informações quando analisadas ao longo do tempo permitem responder questões importantes como tendências de melhoras ou não dos serviços ou produtos vendidos ou ofertados ao público.

Outra proposta deste trabalho é a verificação da viabilidade e implementação de um sistema de recomendação, para que os usuários saibam se seu perfil está muito próximo, do perfil de pessoas que realizaram reclamações, de fornecedores de telefones celulares e *smartphones* escolhidos. Sendo este construído a partir de algoritmos de mineração de dados. Este tópico será abordado mais adiante.

Além da Introdução, este trabalho está dividido nas seguintes seções: a Seção 2 apresenta a revisão bibliográfica e a Seção 3 é constituída pelos trabalhos relacionados; Na Seção 4, são apresentados os Procedimentos Metodológicos que descreve o passo a passo da solução utilizada; Em seguida, são relatados quais análises foram realizadas; e Finalmente, as Considerações Finais sobre o trabalho realizado e a relevância das informações que foram obtidas.

2 REVISÃO BIBLIOGRÁFICA

2.1 Dados Abertos

O conceito de dados abertos, é que são dados que podem ser usados livremente, reutilizados e redistribuídos por qualquer pessoa, no mais se pode ter à exigência de atribuição da fonte dos mesmos, e o compartilhamento pelas mesmas regras. Podem ser de várias fontes, governamentais, pessoais, empresariais. Sendo assim qualquer entidade pública, privada, governamental, pessoa física ou não, pode livremente disponibilizar seus dados.

Visando a transparência principalmente de informações dos órgãos públicos, o Governo Federal Brasileiro tem uma iniciativa de expor a sociedade por meio da internet dados governamentais. Tais dados encontram-se em arquivos eletrônicos, onde qualquer cidadão pode ter acesso e fazer análises, explorar, tornando-se assim não mais alheio às informações que antes eram restritas a poucos. Uma grande quantidade de informação é disponibilizada em dados não-estruturados, semi-estruturados e estruturados, por exemplo, arquivos de extensão pdf, doc, xml, xls, html, csv, json entre outros. Estes dados podem ser livremente manipulados para que seja possível extrair conhecimento. Após esta manipulação, a informação se torna acessível de forma compreensível aos demais cidadãos.

Segundo Machado e Parente de Oliveira(2011), "no âmbito governamental, dados abertos se referem a publicação de dados em formato natural (raw), porém que os tornem acessíveis, prontamente disponíveis para todos e passíveis de reuso".

De acordo com Bico (2012),

> a divulgação de dados governamentais pode ser identificada como um grande passo rumo a um maior envolvimento dos cidadãos na gestão e desenvolvimento da sociedade, uma vez que informação de qualidade sobre as decisões das esferas governamentais poderiam se tornar de amplo conhecimento da população.

Este trabalho utiliza a bases de dados abertos do PROCON e tem como objetivo prover a legibilidade desses dados, conceito explicado em seguida.

2.2 Legibilidade de Dados

Legibilidade de dados compreende o aspecto de tornar facilmente compreensível os dados disponibilizados. No caso deste projeto, os dados abertos que serão manipulados são

os do PROCON, que estão dispostos em arquivos em formato CSV. Neste formato torna-se difícil a visualização ou manipulação e compreensão por parte dos cidadãos leigos, porém interessados em obter informações significativas sobre o mesmo.

Um dos trabalhos realizados sobre o aspecto de legibilidade de dados é o de Legibilidade em Dados Abertos: uma Experiência com os Dados da Câmara Municipal de São Paulo (BICO, 2012).Esse trabalho foi realizado com dados disponibilizados pela Câmera Municipal de São Paulo em formato XML. Por meio de uma aplicação, a população realiza consultas ao assunto de seu interesse. O artigo apresenta uma metodologia utilizada para a construção de uma aplicação, que disponibiliza informações compreensíveis a qualquer usuário. Nesse trabalho, as dificuldades encontradas foram a complexidade do texto contábil e a dificuldade em encontrar metadados, para ajudar a compreensão das informações disponibilizados em arquivos XML.

Outro exemplo de aplicação que trata de obter informações relevantes de conhecimento extraído de bases de dados é o Sistema Estatístico Criminal (SIECRIM) (Silva, 2004). Esse sistema trata da automatização da produção de relatórios estatísticos que auxiliam na análise criminal, e na tomada de decisões para ações estratégicas. Uma característica importante deste sistema é a possibilidade de integração das diferentes bases de dados da Secretaria de Segurança Pública do Estado do Pará. Dentre as dificuldades destacadas no artigo foram: (i) a quantidade massiva de dados na ordem de terabytes, (ii) transformação de dados em informações significativas, (iii) confecção de relatórios em tempo hábil, (iv) necessidade de tratamento de dados, pois os mesmos contêm informações inconsistentes ou ausentes. Além da necessidade de inclusão e exclusão de atributos que tiveram de ser feitas, podendo tornar as informações imprecisas.

É importante destacar as dificuldades encontradas em ambos os trabalhos para que os dados tenham legibilidade. Em relação a este trabalho, pode-se notar que é importante essas iniciativas de trabalhar esses dados, para que os mesmos sejam transformados em informações de modo a gerar conhecimento, e isto não é uma tarefa simples.

2.3 Análise de Dados

2.3.1 DATA WAREHOUSE

Um *Data Warehouse* é um banco de dados, que é uma coleção de informações, bem como sendo um sistema de suporte a decisão. Diferentemente dos bancos tradicionais, que são relacionais, orientados a objeto, em rede ou hierárquico. Pois estes possuem características que os leva a serem utilizados por aplicações de apoio a decisão, além de serem otimizados para possibilitar melhor desempenho na recuperação de dados(SILVA, 2013).

Segundo Ramakrishnan e Gehrke (2000), Os *Data Warehouses* contém dados consolidados de muitas fontes, que tornam-se ricos com sumarização de informações, e cobrem um período de tempo longo. Estes são muito maiores do que outros tipos de bases de dados, pois seus tamanhos variam de vários gigabytes a terabytes. Cargas de trabalho típicas que envolvem, consultas complexas, *ad hoc* e respostas rápidas são de grande importância.

Uma das abordagens de *Data Warehouse* é a *Relational On-Line Analytical Processing* (ROLAP) que segundo Kotidis e Roussopoulos (1998), está emergindo como uma abordagem dominante de data warehouse para apoio à decisão. Com o intuito de melhorar o desempenho de consultas, o ROLAP abordagem baseia-se na seleção, materializando em tabelas de resumo, subconjuntos apropriados de visões agregadas que são então sumarizados para acelerar consultas OLAP.

O esquema estrela é o modelo de dados lógico mais utilizado, composto por apenas dois tipos de tabela: uma tabela de fatos, geralmente posicionada no centro do esquema, e várias dimensões ligadas a essa tabela central (SONG, 2009). Uma tabela de fatos armazena dados conhecidos como medidas, sendo identificada por uma chave primária composta por chaves estrangeiras para todas as dimensões do esquema, mantendo assim um relacionamento com cada uma das dimensões. Uma dimensão armazena atributos que servem como um eixo de análise dos dados, que podem estar organizados em uma hierarquia de atributos.

Um *Data Mart* é visto por (SONG, 2009) como um DW de pequeno porte ou departamental, pois os dados de ambos os repositórios compartilham as mesmas características, isto é, os dados são orientados a assunto, integrados, não voláteis e históricos. Ademais seus dados são organizados em diferentes níveis de agregação. Um *Data Mart* é caracterizado como de pequeno porte, pois seu volume é limitado aos dados de interesse a um departamento, ao invés de atender às necessidades de toda empresa. Outra característica de um *Data Mart* refere-se ao nível de agregação, de forma a agregar os dados a um nível consistente com as necessidades de seus usuários. Armazenar dados agregados, mesmo que

seja em um nível de pequena granularidade, reduz o tempo de resposta no processamento de consultas OLAP, simplifica o entendimento de seu projeto e a sua manutenção.

Devido à grande quantidade de dados oriundos dos arquivos escolhidos para este trabalho, optou-se por criar um *Data Mart*, devido suas características importantes como: armazenar e analisar grandes quantidades de dados. O *Data Mart* criado possibilitará a inserção de novos dados, a qualquer momento que sejam disponibilizados, de forma a se utilizar um processo incremental, possibilitando assim realizar novas analises, junto com os novos dados inseridos.

A utilização de OLAP no contexto deste projeto ocorre através das consultas realizadas no banco ROLAP criado para tal propósito, pois as mesmas buscam através do uso de consultas realizadas em linguagem SQL ou de ferramentas de inteligência de negócios, extrair informações significativas dos dados presentes no banco e apresentá-las neste trabalho de forma compreensível, através de gráficos e outras formas que proporcionem a qualquer usuário leigo em sistemas de informação, compreender as informações extraídas dos arquivos. Pois a intenção é que as informações sejam úteis ao público, agregando assim grande valor a mesma.

Segundo Codd, Codd e Saley (1993):

> As ferramentas de consultas/relatórios e as planilhas eletrônicas têm sido extremamente limitadas nas formas pelas quais os dados (já recuperados do SGBD) podem ser agregados, resumidos, consolidados, somados, visualizados e analisados. A carência mais notada tem sido a capacidade para consolidar, visualizar e analisar dados de acordo com múltiplas dimensões, de maneira que faça sentido para um ou mais analistas específicos em um determinado ponto no tempo. Este requisito é chamado "análise de dados multidimensionais". Talvez um melhor e mais genérico nome para este tipo de funcionalidade é *Online Analytical Processing* (OLAP), em que a análise de dados multidimensionais é apenas uma de suas características.

No ano de 1995, *THE OLAP COUNCIL* (1995), um conselho de padronização da tecnologia realizou uma publicação da conceitualização do termo da seguinte forma:

> *Online Analytical Processing* (OLAP) é uma categoria de tecnologia de software que possibilita que os analistas, gerentes e executivos tenham entendimento sobre os dados de forma rápida, consistente, e com acesso interativo a uma ampla variedade de visões possíveis de informações que foram transformadas a partir de dados brutos para refletir a dimensionalidade real da empresa como entendida pelo usuário.

2.3.2 Mineração de Dados

Mineração de dados é o termo que se popularizou para denominar o processo de descoberta de conhecimento em bases de dados. Trata-se da utilização de ferramentas computacionais a fim de descobrir informações valiosas, potencialmente úteis, descritas na forma de padrões, a partir dos volumes de dados que estão sendo coletados e armazenados pelas organizações atualmente (FAYYAD et al., 1996).

Segundo Rezende (2003), existem etapas anteriores ao processo de mineração de dados são estas: Pré-processamento, Extração de Padrões e Pós-processamento, segundo o mesmo referem-se ao conhecimento do domínio e identificação do problema, e uma fase depois ao processo, sendo esta a de utilização do conhecimento que fora adquirido.

Segundo Silva (2004), as etapas do KDD (*Knowledge Discovery in Database*) que tratam da descoberta de conhecimento em bases de dados, são definidas desta maneira:

> "O processo de descoberta de conhecimento em banco de dados é feito de forma interativa, iterativa, cognitiva e exploratória, nos quais estão envolvidos vários passos e as decisões são feitas pelo analista do domínio ou analista de dados."

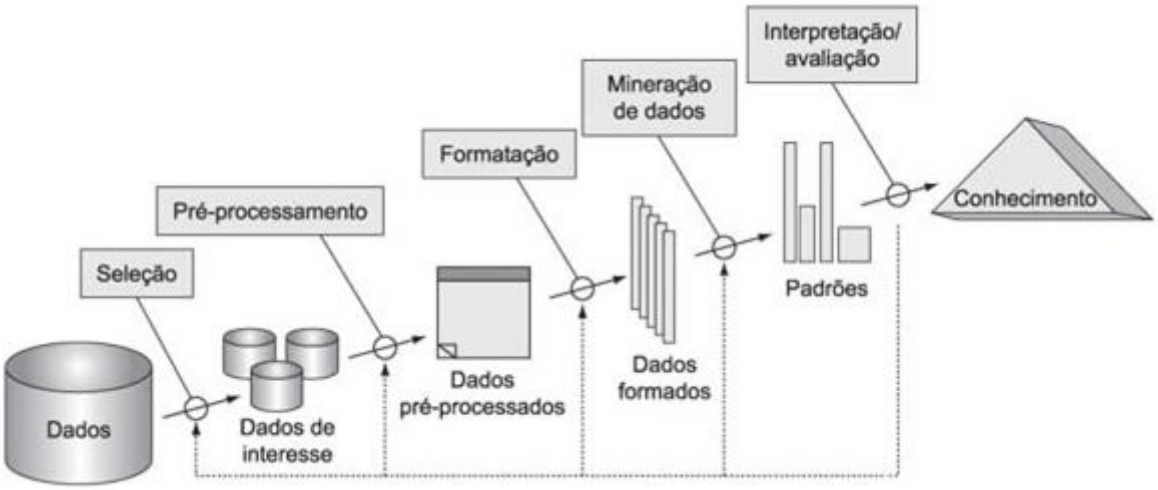
Segundo o autor segue então as descrições das etapas do processo de KDD.

1. Definição de qual o conhecimento se quer descobrir entendendo-se o domínio, e o tipo de decisão que conhecimento proporcionará em termos de melhora;
2. Seleção: define-se como a criação do conjunto de dados alvo, no qual o foco será direcionado a um subconjunto do mesmo a ser trabalhado;
3. Limpeza de dados e pré-processamento: nesta fase são realizadas: remoção de ruídos se houver necessidade, coleta da informação para modelagem ou estimar ruído, escolha de estratégia para manipular dados ausentes, formatação dos mesmos para que tornem-se adequados à mineração;
4. Transformação: consistindo na redução de dados e projeção, encontrar características que serão úteis para representação dos dados, dependendo do objetivo da tarefa, diminuindo assim a quantidade de variáveis e/ou instâncias que serão consideradas para o conjunto de dados, e também o enriquecimento semântico das informações;
5. Mineração de dados:etapa de selecionar métodos a serem utilizados para localização de padrões nos dados, seguida de busca por padrões de interesse particular, buscando o melhor ajuste de algoritmos de mineração para a tarefa;

6. Interpretação/Avaliação são retornos aos passos 1 ao 6, para posteriormente seguindo-se com a iteração;
7. Conhecimento: etapa final que consiste em utilizar, implantar, documentar e reportar a quem é de interesse.

Na Figura 1 é perfeitamente ilustrada as etapas do processo de KDD, visualizando o sentido em que ocorrem cada fase do processo.

Figura 1- Etapas do Processo de KDD.



Fonte: Fayyad et al.(1996).

Segundo Ferreira (2005), o último objetivo do KDD é conseguir extrair conhecimento inteligível e utilizá-lo para dar apoio a decisões, e não apenas simplesmente encontrar padrões e relações na enorme quantidade de dados existentes em bases de dados.

O processo de minerar dados é formado por um conjunto de técnicas para descoberta de conhecimento a partir de grandes bases de dados. Tais técnicas baseiam-se em modelos capazes de sumarizar dados, extrair novos conhecimentos ou realizar predições. A fase de Mineração de Dados como processo KDD pode ser substituída pelo tipo de análise OLAP.

Existem muitos trabalhos relacionados a estes aspectos, como por exemplo: Modelo de Mineração de Dados para classificação de clientes em telecomunicações (FERREIRA, 2005), cujo foco é a utilização de técnicas de mineração de dados para obtenção de informações relevantes sobre os clientes de empresas de telefonia, a fim de que estas informações venham a dar um norteamento sobre o que deve ser feito para a manutenção de seus clientes, pois isto é crucial em um mercado competitivo.

No artigo Minerando e Caracterizando Dados de Currículos Lattes (DIGIAMPIETRI, 2012), é apresentada a importância da mineração de dados, para a construção de um banco de dados com mais de um milhão de currículos contendo informações sobre pesquisadores, proporcionando produção e análises de redes sociais destes profissionais. O autor destaca todo processo realizado para coletar e processar os dados, de forma a estarem prontos para serem inseridos em um banco de dados relacional, e assim então realizar as consultas que retornarão as informações buscadas. Um problema enfrentado nesse trabalho foi a dificuldade em obter os dados junto ao CNPq.

Outro trabalho é o artigo sobre Legibilidade e Mineração de Dados na Web para Inteligência Competitiva (DE ALMEIDA, 2004). O uso de mineração de dados na web mostra um grande potencial antes não explorado, onde esse trabalho traz em detalhes formas do uso de mineração de dados utilizado na web para se obter informações, que mais tarde serão utilizadas por empresas do mercado.

Por fim, o artigo sobre Mineração de Dados Educacionais: Oportunidades para o Brasil (BAKER, et. al., 2011), vem mostrar importância deste tipo de análise sobre dados. Neste caso sobre dados gerados pelos alunos ao utilizarem plataformas de softwares educacionais. Através destes softwares se obtém grande conhecimento sobre os alunos, sendo possível até mesmo traçar o perfil do comportamento dos alunos. O trabalho aborda o potencial impacto da EDM (Mineração de Dados Educacionais) na melhora da qualidade dos cursos na modalidade educação a distância, que vêm recebendo incentivo governamental e um crescente número de alunos matriculados.

2.3.3 Sistemas de Recomendação

A grande quantidade de informações disponíveis na Web, bem como a interação dos usuários com os sistemas, tem favorecido o surgimento de uma extensa classe de aplicações que envolvem predizer respostas dos usuários referentes a um conjunto de opções. Esses sistemas são chamados de Sistemas de Recomendações, que constitui outra área importante de Mineração dados. Recomendação de artigos de notícias para leitores on-line, sugestão de produtos, filmes a usuários são exemplos de aplicações que utilizam sistemas de recomendação.

Os sistemas de recomendação podem ser classificados basicamente em dois grupos:

- Sistemas baseados em conteúdo: utilizam propriedades dos itens a serem recomendados;
- Sistemas com filtragem colaborativa: similaridade entre usuário e/ou itens são utilizados;

O artigo sobre uso de técnicas de recomendação em um sistema para apoio à aprendizagem colaborativa(LICHTNOW; GARIN; PALAZZO, 2006), apresenta o SisRecCol – Sistema de Recomendação para Apoio à Colaboração. O mesmo apresenta ferramentas que tem por finalidade apoiar o processo de aprendizagem colaborativa. Este, detalha os módulos existentes no sistema e as metodologias utilizadas para apoiar uma recomendação. A partir da identificação dos perfis de usuários, o sistema realiza recomendações, utilizando filtragem baseada em conteúdo e filtragem colaborativa. O Sistema realiza as recomendações a partir da análise do comportamento dos usuários em um Web Chat e do acesso a uma biblioteca digital. Ao final, o autor mostra que foram realizados testes, com a utilização do sistema por alunos de um determinado curso de tecnologia, e conseqüentemente tem como *feedback*, resultados significativos para a realização de mudanças e melhorias. Sendo assim foi possível identificar, qual das duas técnicas empregadas obteve melhor resultado no que diz respeito às recomendações feitas aos alunos.

O trabalho sobre Recomendação de Objetos de Aprendizagem Empregando Filtragem Colaborativa e Competências (CAZELLA et al., 2009), propõe um sistema computacional para fazer a recomendação personalizada de objetos de aprendizagem (OA), de acordo com as predileções (“gostos” por determinados objetos de aprendizagem) de cada aluno, o mesmo utiliza filtragem colaborativa e competências. O sistema permite que alunos recebam a recomendação de forma automática conforme seus interesses, de acordo com as competências que devem ser desenvolvidas dentro de um plano de aula.

A construção de um Sistema de Recomendação é baseado em uma matriz de utilidade de acordo com Ullman (2012, p.306). Um exemplo de matriz de utilidade é apresentado na Figura 2. Imagine que em um sistema de recomendação existem duas classes de entidades: usuários e itens. Os usuários têm preferências por determinados itens, e estas preferências devem ser mapeadas. Para cada par (usuário, item), o valor que está associado ao par,representa o que se sabe sobre o grau de preferência que o usuário tem para esse item.

Na figura abaixo um exemplo de uma matriz de utilidade, a qual utiliza uma pontuação de 1-5, dada pelos usuários A, B, C e D aos respectivos filmes das colunas. Caracterizando assim sua nota para os mesmos.

Figura 2 - Matriz de Utilidade representando pontuação de filmes em escala de 1-5 por usuários A, B, C e D.

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Fonte: Ullman (2012. p. 306)

Neste trabalho, através de uma matriz, pode-se realizar uma análise qualitativa sobre as reclamações referentes aos dados do PROCON. Inicialmente, o perfil de reclamações de cada empresa foi construído baseado em propriedades presentes nas reclamações. Sendo assim, a aplicação proposta poderia receber como entrada um perfil de usuário e assim, recomendar de qual empresa este usuário não deverá obter um produto ou serviço, o perfil de quem reclama na empresa é similar ao do usuário.

Neste trabalho foi analisada a viabilidade de construção desse sistema de recomendação explicado acima. Este assunto será discutido na sessão de procedimentos metodológicos.

3 TRABALHOS RELACIONADOS

A seguir, são apresentados os principais trabalhos relacionados ao contexto deste trabalho.

O artigo intitulado Proposta de Data Mart para Análise de Reclamações Realizadas em PROCONs (PEREIRA; ALMEIDA FILHO; SOUZA, 2013), utilizou dados obtidos do portal de dados abertos. O intuito do trabalho foi de propor um modelo de *Data Mart* para os dados do PROCON, para possibilitar análises. Porém, as análises realizadas se limitam a dados de um trimestre. Para analisar, os autores utilizaram a ferramenta *Pentaho Business Inteligence*, a mesma utilizada neste trabalho. O que distingue este trabalho e o artigo em questão diz respeito a quantidade de dados manipulados, a proposta de um modelo ROLAP para o *Data Mart* construído, e a utilização de toda a base de reclamações do

PROCON de 2012. Além disso, (PEREIRA; ALEMIDA FILHO; SOUZA, 2013) não estudaram a viabilidade de construção de um Sistema de Recomendação.

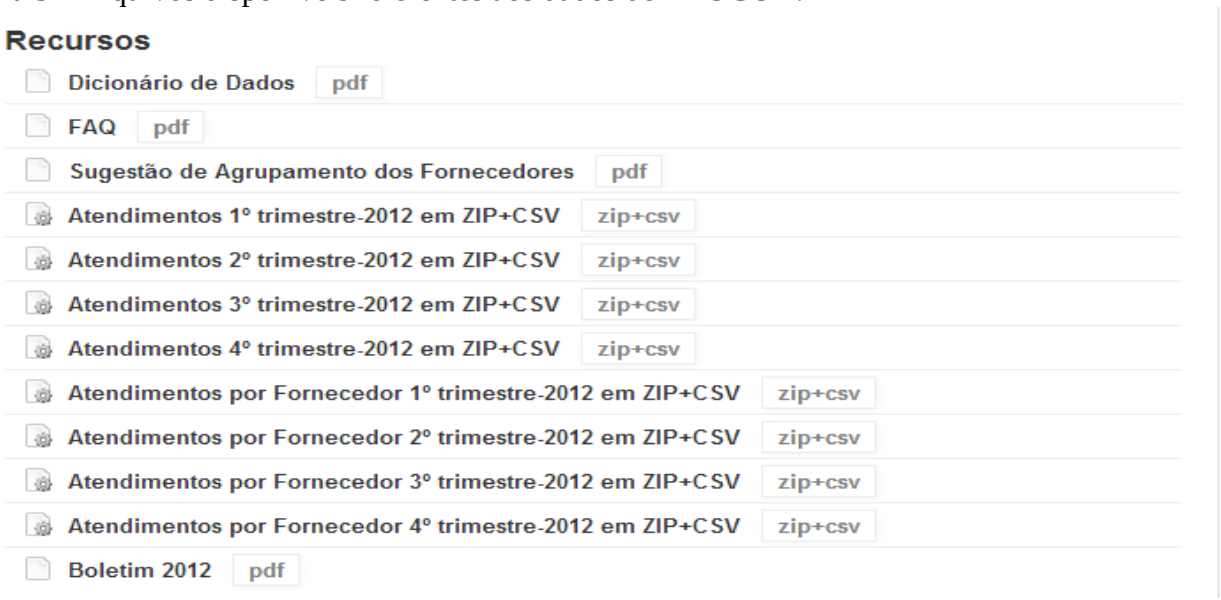
Outro trabalho relacionado a ser considerado é uma Proposta de um *Data Mart* para Avaliação de Empresas Usuárias do *Twitter* através das Mensagens Postadas pelos Clientes(MAGALHÃES et al., 2012).Tal trabalho apresenta um *Data Mart* que permite a análise da reputação das empresas que utilizam o *Twitter* com o objetivo de fornecer apoio/informações para a tomada de decisão de seus futuros consumidores e fornecedores, além de informações para definir estratégias competitivas delas mesmas, em relação a seus concorrentes. Esse trabalho utiliza um *Data Mart* e a ferramenta *Pentaho* para a fase de Extração, Transformação e Carga da base de dados e ainda nas análises realizadas via OLAP. O trabalho de (MAGALHÃES et al., 2012) difere do realizado neste pelos dados que foram manipulados. Porém, é semelhante o intuito de utilizar-se de um *Data Mart*, para realizar análises, e através das mesmas obter informações.

4 PROCEDIMENTOS METODOLÓGICOS/RELATO GERAL DO DESENVOLVIMENTO

Essa seção apresenta os procedimentos metodológicos adotados neste trabalho. As ferramentas utilizadas foram o SGBD PostgresSQL, para gerenciar, armazenar a base de dados; o SQL Power Architect (uma ferramenta visual de design para o banco de dados PostgresSQL), para modelagem do banco de dados; e o framework *Pentaho* para criar, administrar e realizar todas as operações de ETL, processo no qual os dados foram extraídos dos arquivos fontes e armazenados em um *Data Mart*, e OLAP, para que os dados armazenados sejam recuperados e analisados de forma rápida e fácil, possibilitando-se extrair conhecimento, tendências, realizar comparações e destacar problemas.

Inicialmente foi realizada a escolha de quais dados, dos que estão disponíveis no portal de dados abertos do governo federal, seriam utilizados neste projeto. No site constam dois modelos de arquivos CSV disponíveis, referentes aos dados do PROCON, que são atendimentos referentes aos trimestres de 2012 e atendimentos por fornecedor dos trimestres de 2012, conforme figura 4 que mostra a forma em que os arquivos estão dispostos no portal.

Figura 3 - Arquivos disponíveis referentes aos dados do PROCON.



Fonte: Portal Dados Abertos (2013).

Neste projeto optou-se por utilizar os arquivos referentes a atendimentos por fornecedor, já que mesmos mantém o mesmo padrão dos arquivos de atendimento, com o acréscimo de informações pertencentes aos fornecedores, de serviços ou produtos que tiveram reclamações registradas. Após a coleta desses dados, foi utilizado o processo de KDD(*Knowledge Discovery in Database*) explicado anteriormente que compreende todas as fases de limpeza dos dados, bem como análise e interpretação dos resultados.

No quadro abaixo visualiza-se os campos existentes e suas descrições nos arquivos CSV.

Campo	**Descrição**
AnoAtendimento	Ano do Atendimento
TrimestreAtendimento	Trimestre do Atendimento
MesAtendimento	Mês do Atendimento
DataAtendimento	Data do Atendimento
CodigoRegiao	Código da Região
Regiao	Região
UF	Unidade Federativa
CodigoTipoAtendimento	Código do Tipo de Atendimento
DescricaoTipoAtendimento	Descricao do Tipo do Atendimento
GrupoAssunto	Grupo do Assunto
CodigoAssunto	Código do Assunto
DescricaoAssunto	Descrição do Assunto

GrupoProblema	Grupo do Problema
DescricaoProblema	Descrição do Problema
CodigoProblema	Código do Problema
SexoConsumidor	Sexo do Consumidor
FaixaEtariaConsumidor	Faixa Etária do Consumidor
CEPConsumidor	CEP do Consumidor
TipoFornecedor	Tipo do Fornecedor
RazaoSocialSindec	Razão Social Sindec
NomeFantasiaSindec	Nome Fantasia Sindec
CNPJ	Cadastro Nacional de Pessoa Jurídica
RadicalCNPJ	Radical CNPJ
RazaoSocialRFB	Razão Social RFB
NomeFantasiaRFB	Nome Fantasia RFB
CodigoCNAEPrincipal	Código CNAE Principal
DescricaoCNAEPrincipal	Descrição CNAE Principal

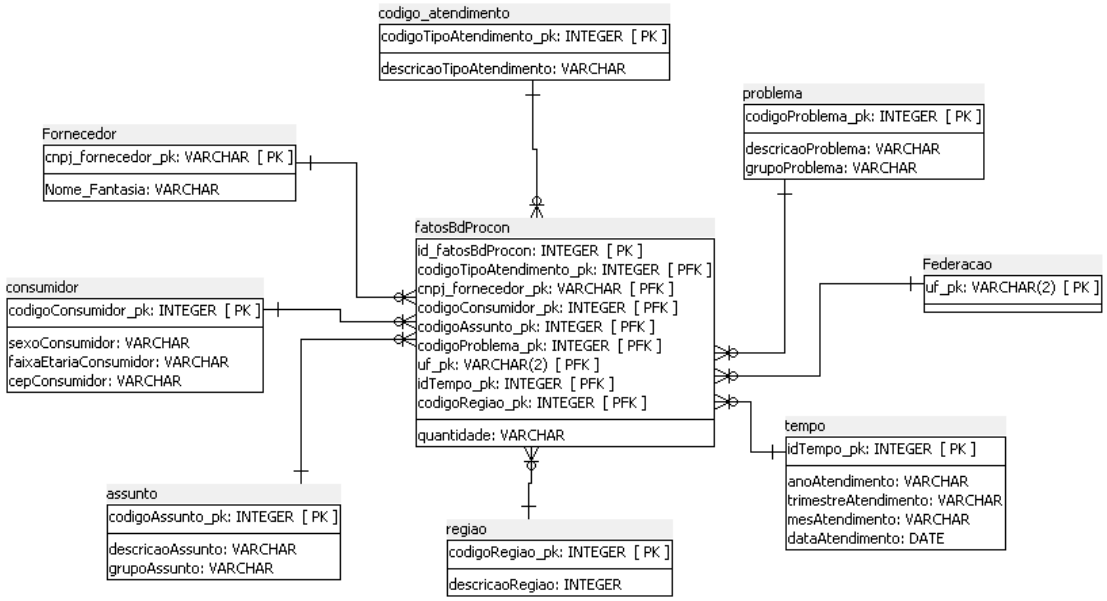
A primeira fase do processo consiste em realizar a seleção de quais registros dos arquivos CSV referente a reclamações do PROCON, que foram obtidos do portal de dados fase de pré-processamento considerou cada registro do CSV, como uma reclamação distinta e os arquivos CSV foram divididos manualmente em partes para viabilizar o processamento de carga no banco de dados, pois devido a capacidade computacional da máquina utilizada, fez-se estão necessário dividí-los em arquivos com média de 69.000 linhas.

Na fase de formatação foi escolhido um modelo de banco de dados que fosse mais adequado as consultas a serem realizadas. Dessa forma, optou-se pelo modelo ROLAP(*Relational On Line Analytical Processing*).O esquema utilizado na modelagem do banco de dados foi o modelo Estrela, pois apresenta melhor desempenho no processamento de consultas OLAP, evitando a realização de várias junções como ocorre no modelo Flocos de Neve (CALDEIRA, 2012) .

> "O modelo *Snow Flake* (floco de neve) é uma variação do modelo estrela. Esse possui a mesma abordagem de colocar o fato ao centro e as dimensões ao seu redor. Contudo, sua abordagem separa as hierarquias das dimensões em tabelas diferentes, variantes da dimensão principal. Este modelo é resultado da terceira forma normal nas dimensões, evitando a redundância de valores textuais em uma tabela e deixando mais visível as hierarquias. Porém, esta abordagem pode deixar o modelo bastante poluído à medida que aumentam as dimensões presentes no projeto. Com isso, ao invés de facilitar a visualização dos dados, há uma dificuldade de identificar as dimensões principais e as hierarquias variantes delas." (MACHADO, 2006).

No banco de dados foi criada uma tabela de fatos com uma medida quantitativa, e as dimensões do banco como mostrado na Figura 4.

Figura 4 - Modelagem de Banco ROLAP – Modelo Estrela, com a ferramenta SQL Power Architect.



Fonte: elaborada pelo autor.

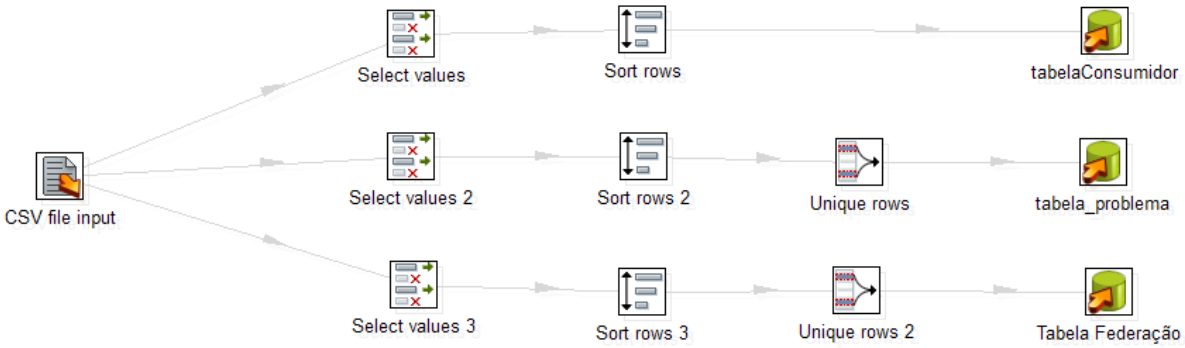
O servidor de banco de dados utilizado para este trabalho foi o PostgresSQL, pois o mesmo possuí a qualidade de ser um software livre, sendo bastante robusto, possuindo bibliotecas para numerosas linguagens, e uma API que permite a qualquer aplicação que suporta este tipo de interface acessar as bases de dados. O framework *Pentaho DataIntegration* foi utilizado para realizar o processo de povoamento do banco de dados ou ainda Extração, Transformação e Carga. Ressaltando ainda, que não foram cadastradas informações duplicadas, proporcionando assim que o banco não tenha informações desnecessárias, já que os arquivos somam juntos mais de 1.500.000 (um milhão e meio) de linhas em informações.

Inicialmente é selecionado o arquivo CSV que será utilizado como fonte dos dados para o *Data Mart* através do objeto CSV *file input*. No passo seguinte para realizar a população de uma tabela, utiliza-se um objeto *select values*, para que se possa projetar quais colunas do arquivo CSV serão inseridos na tabela escolhida, posteriormente é feita a ordenação dos dados que serão inseridos por meio de um objeto *sort rows*. Caso exista a necessidade de filtragem de dados, de modo que não sejam inseridos dados duplicados, como por exemplo chaves primárias, faz-se então necessário a utilização de um objeto *unique rows*.

Por fim um *table* é adicionado para receber os dados que serão processados nas etapas anteriores, sendo que o mesmo irá popular a tabela real criada no *Data Mart.*

Na Figura 5 é apresentado o *Spoon* que é uma ferramenta do *Pentaho Data Integration* utilizada para povoamento do *Data Mart* e os passos utilizados para popular algumas tabelas do *Data Mart*.

Figura 5–Pentaho- Spoon.

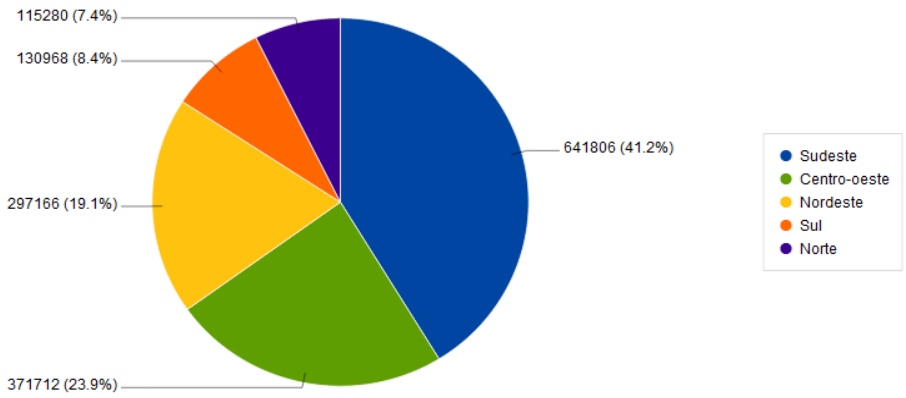


Fonte: Spoon – Pentaho Data Integration.

Na fase de análise o autor utilizou o *Pentaho Business Analytics.* Por meio desse framework, as consultas mostradas na Seção de análises foram realizadas, a fim obter informações quantitativas sobre as reclamações feitas por homens e mulheres, entre outros exemplos.

O framework foi utilizado no processamento das consultas OLAP, e por meio dele foi possível gerar os gráficos, que auxiliaram na fase de interpretação dos resultados. Logo abaixo a Figura 6 mostra um gráfico gerado a partir de consulta realizada com o *Pentaho*.

Figura 6 - Gráfico apresenta o percentual e a quantidade de reclamações registradas por região.



Fonte: elaborada pelo autor. Pentaho Business Analytics.

Quanto a fase de análise optou-se por 2 formas, como já explicado anteriormente. Uma por OLAP que será descrita na próxima seção e outra via Mineração de Dados. Com relação a Mineração de Dados, optou-se pela construção de um sistema de recomendação, a fim de descobrir se existe para cada fornecedor de produto, um perfil de cliente insatisfeito e que reclama. Dessa forma, um usuário interessado em comprar de determinado fornecedor, pode utilizar o sistema de recomendação e saber se ele apresenta o mesmo perfil dos clientes insatisfeitos com aquele fornecedor. Para isso foi construída a matriz de utilidade, que é base do sistema de recomendação. Os dados para a construção da matriz foram obtidos utilizando consultas em SQL no *Data Mart* criado. Como os fornecedores que possuem cadastros no PROCON, são inúmeros, então neste trabalho optou-se por escolher alguns fornecedores fabricantes de telefones celulares e *smartphones*.

Os fabricantes escolhidos foram: Samsung, LG, Nokia, Sony, Motorola, ZTE e Alcatel. A matriz de utilidade construída tem como informação, a quantidade de reclamações que cada um destes fabricantes possui registradas. Os itens que traçam o perfil dos usuários que reclamaram são: sexo, faixa etária e região do País na qual residem.

Em seguida, verificou-se a viabilidade na utilização dos dados para cálculo de similaridade e estabelecimento de um perfil de clientes que reclamam de determinado fornecedor. No entanto, observou-se que seria inviável construir esses perfis, já que para cada característica utilizada para formar os perfis (sexo, faixa etária, região), a diferença entre os percentuais de reclamações envolvendo todos os fornecedores para um mesmo valor de característica foi irrelevante. Dessa forma, o percentual de clientes do sexo masculino que reclamam do fornecedor SAMSUNG está muito próximo ao percentual de clientes desse mesmo sexo que reclamam de outros fornecedores, por exemplo. Isto pode ser observado na Figura 7 que exibe a matriz de utilidade construída. Neste caso, a recomendação se tornaria

imprecisa, inviabilizando a construção do sistema de recomendação. O autor não utilizou outros fornecedores pela inviabilidade de testar para todos os fornecedores da base de dados, quais os que forneceriam uma matriz de utilidade bem caracterizada nos perfis de consumidores insatisfeitos.

Figura 7–Matriz de utilidade construída.

	PORCENTAGEM		FAIXA ETARIA								NO	NE	SE	SU	CE
	MAS	FEM	ATE 20	21 A 30	31 A 40	41 A 50	51 A 60	61 A 70	M 70	NI	1	2	3	4	5
SAMSUNG	47,46	52,54	3,82	28,44	28,60	18,60	9,95	5,03	1,41	4,14	12,57	44,00	23,42	3,20	16,82
LG	50,64	49,36	2,79	27,33	29,23	19,55	11,28	4,11	1,02	4,69	13,71	43,65	21,58	2,39	18,66
NOKIA	49,22	50,78	3,29	23,51	29,15	20,22	12,28	5,38	0,94	5,22	10,40	57,78	17,03	4,70	10,08
SONY	45,05	54,95	2,83	29,72	29,72	20,75	8,96	2,59	1,18	4,25	12,97	50,00	11,56	4,25	21,23
MOTOROLA	48,43	51,57	4,13	30,45	28,31	18,25	7,86	3,13	0,87	7,00	11,93	36,98	24,05	4,33	22,72
ZTE	39,59	60,41	3,81	21,11	24,63	24,34	12,90	4,40	1,47	7,33	7,04	50,44	18,48	2,64	21,41
ALCATEL	45,35	54,65	1,16	15,12	33,72	22,09	13,95	4,65	2,33	6,98	12,79	56,98	11,63	0,00	18,60

Fonte: elaborada pelo autor.

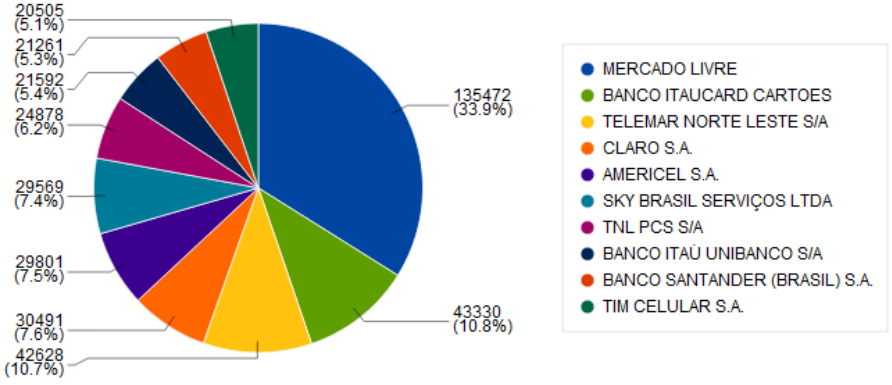
5 ANÁLISES REALIZADAS

Nesta sessão são apresentadas as análises realizadas utilizando-se o *Pentaho Business Analytics*, A ferramenta mostra-se extremamente eficiente para a construção de tabelas e gráficos, dando credibilidade as informações geradas, pois a mesma é muito robusta e bastante utilizada no mercado de Tecnologia da Informação e no meio acadêmico.

No tocante aos dados utilizados um problema enfrentado é falta de padronização, quando em relação por exemplo: o nome dos fornecedores nos quais foram registradas reclamações.

A seguir a Figura 8 mostra os 10 principais fornecedores que mais receberam registro de reclamações, exibindo-se o percentual e a quantidade. Com exceção do Mercado Livre, observa-se que entre os demais fornecedores, não existe uma margem percentual muito grande de diferença quanto ao número de reclamações. Por meio desse gráfico, é possível extrair a informação de quais são as empresas que deixam o consumidor mais insatisfeito.

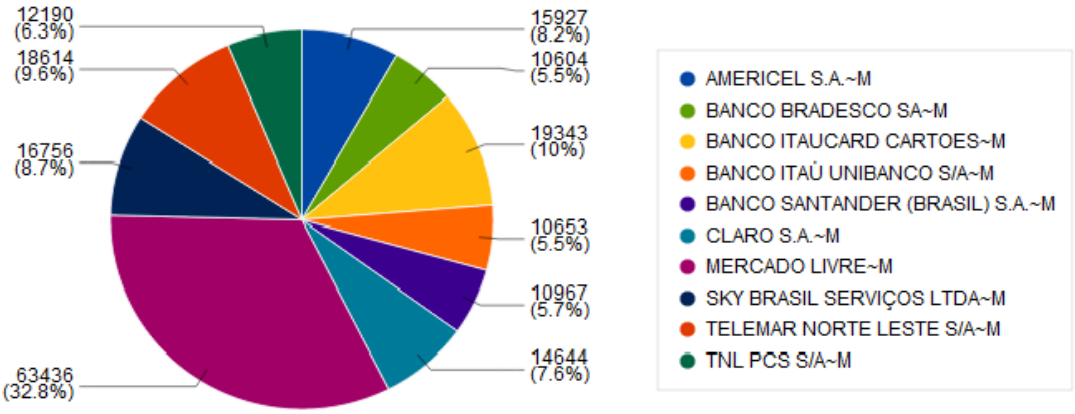
Figura 8 - mostra os 10 principais fornecedores que mais receberam registro de reclamações.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 9 mostra-se os 10 fornecedores que mais tiveram registro de reclamações, junto ao PROCON, realizadas por homens em todas as faixas etárias e regiões.

Figura 9 - Os 10 Fornecedores de quem os homens mais registraram reclamações.

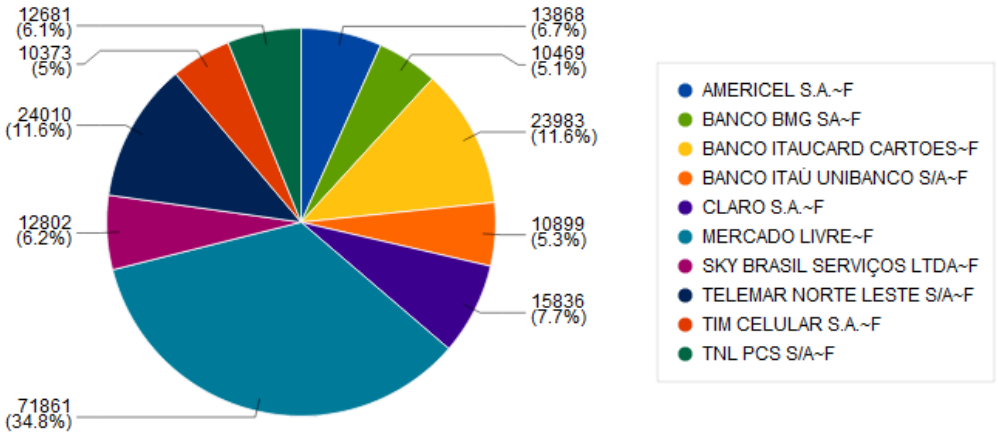


Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 10 é possível visualizar as reclamações registradas pelas mulheres, possibilitando fazer comparações com a figura anterior, de modo a verificar de quais serviços ambos sexos têm semelhanças ou diferenças no registro de reclamações.

Observa-se que tanto homens quanto mulheres, apresentam perfil semelhante quanto ao número de reclamações, referentes as principais empresas que são reclamadas. Como Mercado Livre, empresas de serviços bancários e de telefonia.

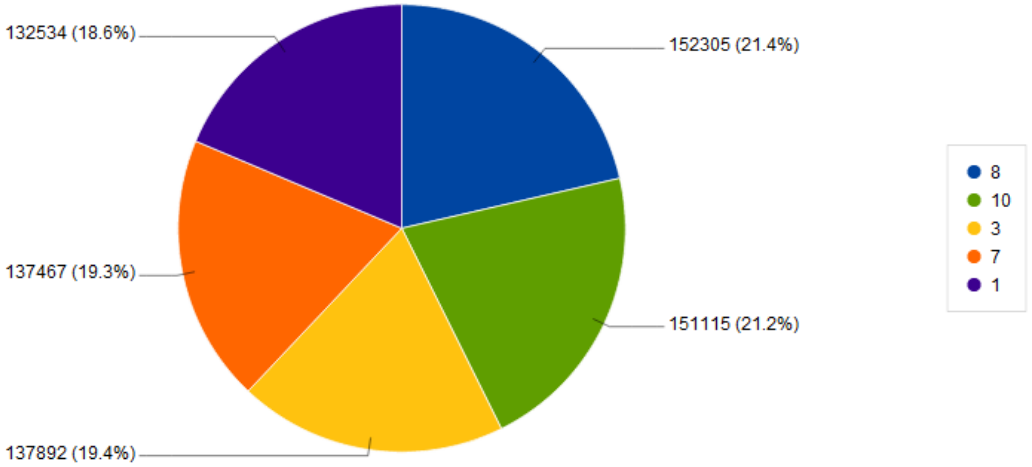
Figura 10 - Os 10 Fornecedores de quem as mulheres mais registraram reclamações.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 11 – Observa-se os 5 meses do ano que mais registraram reclamações junto ao PROCON. Através do gráfico, pode-se que não existiu grande variação percentual entre os mesmos. E que o número de reclamações por mês se mantém quase o mesmo (em média) ao longo dos meses.

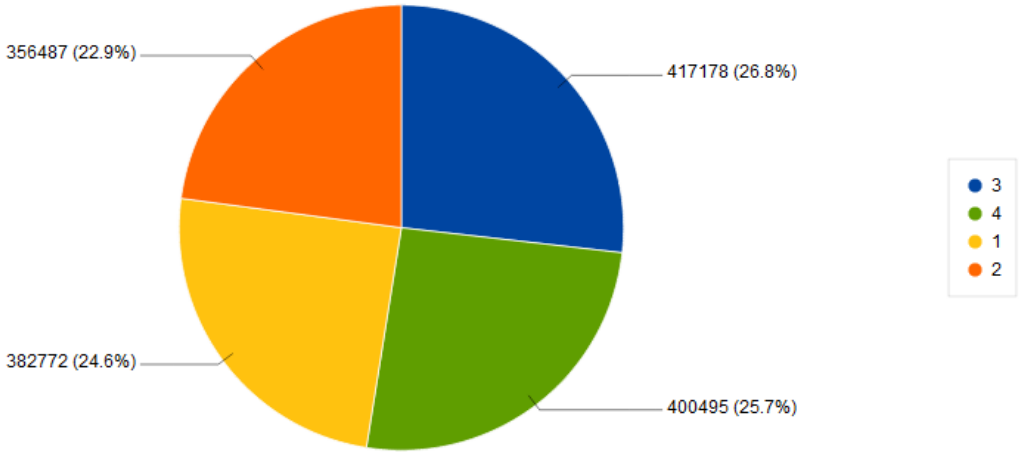
Figura 11 - Os 5 meses que registraram maior número de reclamações.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 12 é possível visualizar um crescimento no número de reclamações, nos dois últimos trimestres, em relação aos dois primeiros do ano, principalmente no terceiro. No entanto, é esperado que por trimestre não ocorra muitas variações quanto ao número de reclamações.

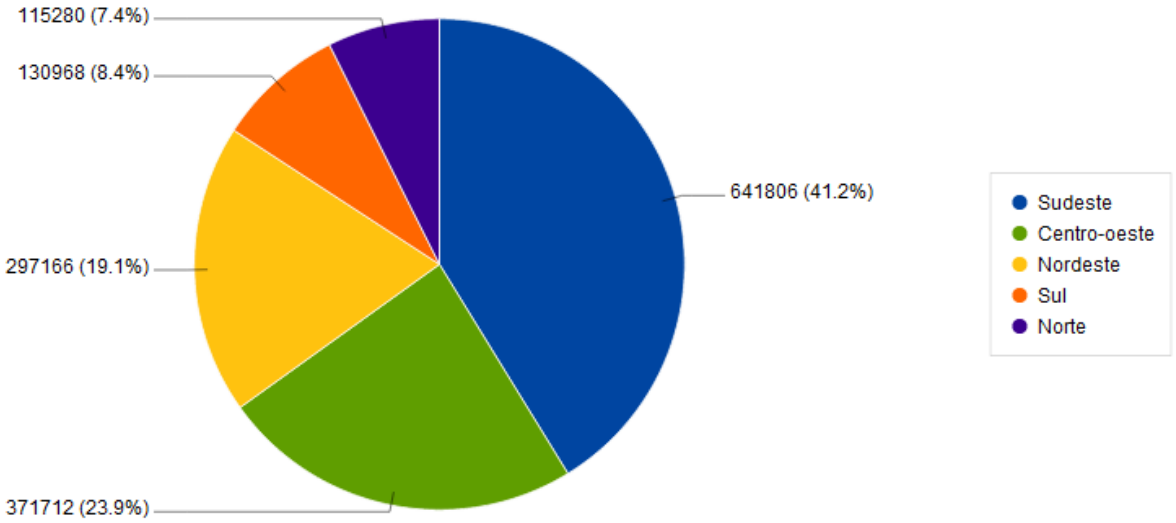
Figura 12 - Percentual e quantidade de reclamações registradas por trimestres.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 13 é possível visualizar que as regiões, onde se encontram os grandes centros urbanos no País, as regiões Sudeste e Centro-Oeste, juntas são responsáveis por mais de 65% do total geral de reclamações registradas pelo PROCON.

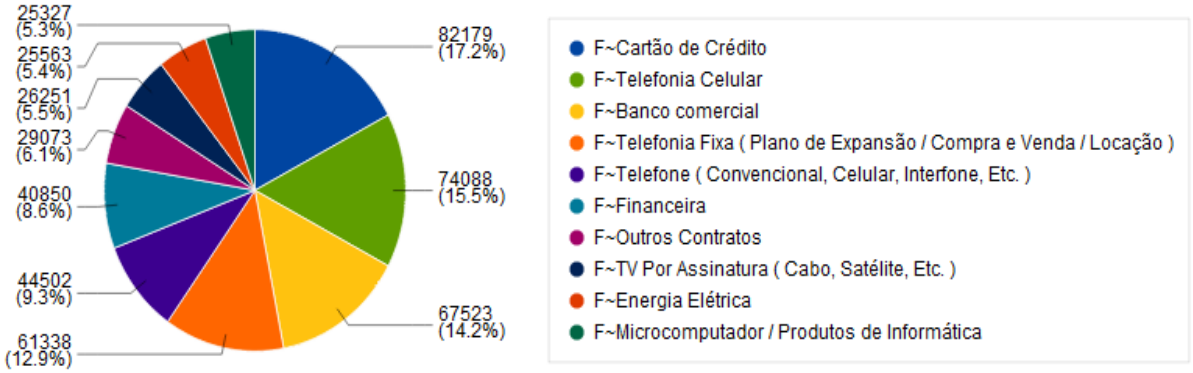
Figura 13 - Percentual e quantidade de reclamações por região do País.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 14 – O gráfico mostra que 4 assuntos predominam entre as reclamações feitas por mulheres, tendo como a principal as alusivas a cartão de crédito. Perceba que essa informação dos principais assuntos é semelhante a gerada na Figura 12 que apresenta os fornecedores os quais as mulheres mais reclamam.

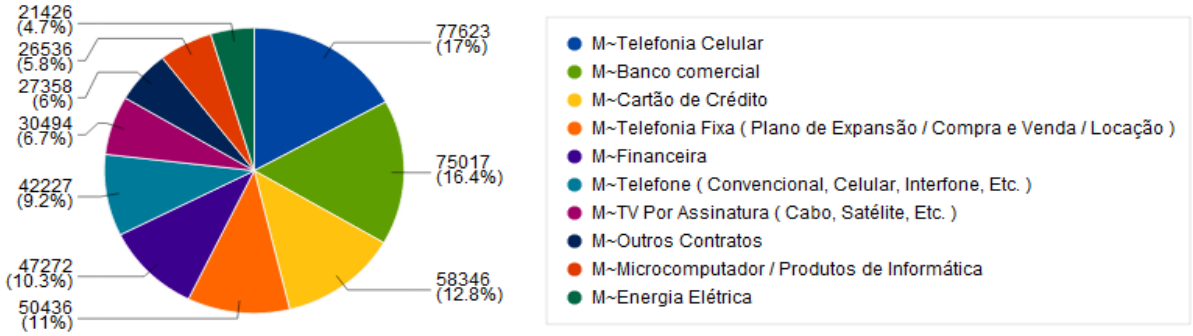
Figura 14- 10 principais assuntos reclamados pelas mulheres.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 15 – No gráfico pode-se perceber que existem 6 assuntos com maior incidência de reclamações feitas por homens, os mesmos junto contabilizam mais de 76% do registro das reclamações, tendo como principal assunto a telefonia celular. Semelhante ao que foi observado na Figura 9.

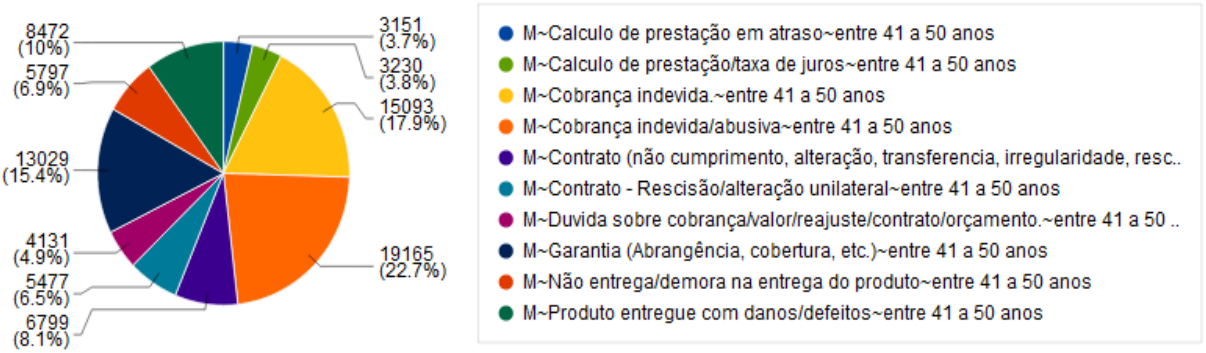
Figura 15 - 10 principais assuntos reclamados pelos homens.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 16 – Mostra os 10 principais problemas registrados por homens, na faixa etária de 41 a 50 anos. Nesta análise destacam-se dos demais problemas: cobrança indevida ou abusiva, cobrança indevida, garantia (abrangência, cobertura, etc.).

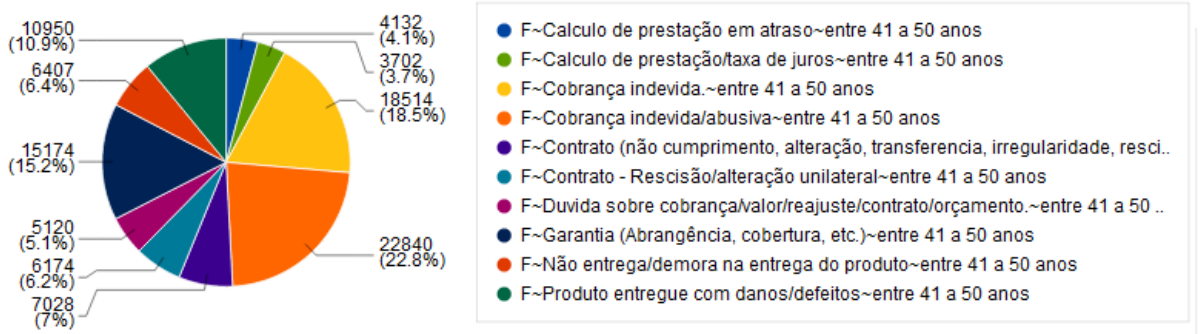
Figura 16 - 10 principais problemas registrados por homens na faixa etária entre 41 a 50 anos.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 17 – Pode-se visualizar que na faixa etária entre 41 a 50 anos de idade, tanto o perfil dos homens visto na figura anterior, quanto as mulheres tem os mesmos 2 problemas dominantes.

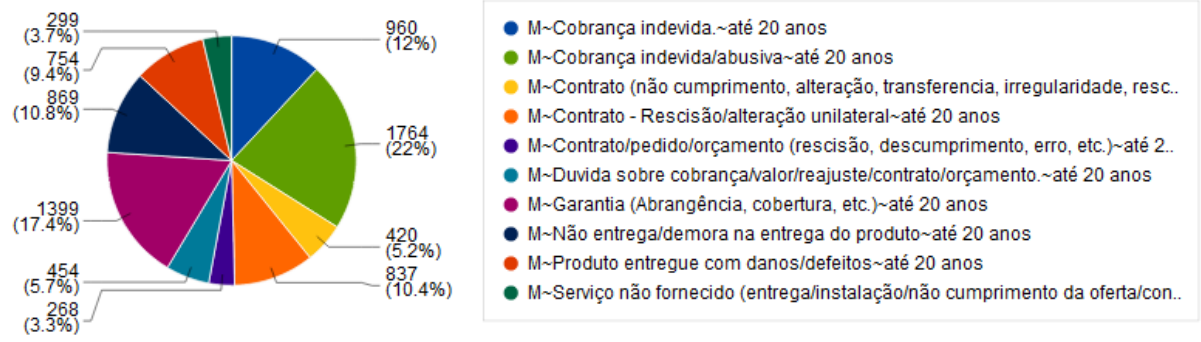
Figura 17–10 principais problemas registrados por mulheres na faixa etária entre 41 a 50 anos.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 18- Podemos ver quais os 10 problemas que jovens de até 20 anos, do sexo masculino, realizam registro de reclamações junto ao PROCON. Verifica-se então que os mesmo problemas registrados com os adultos, também são predominantes entre os jovens.

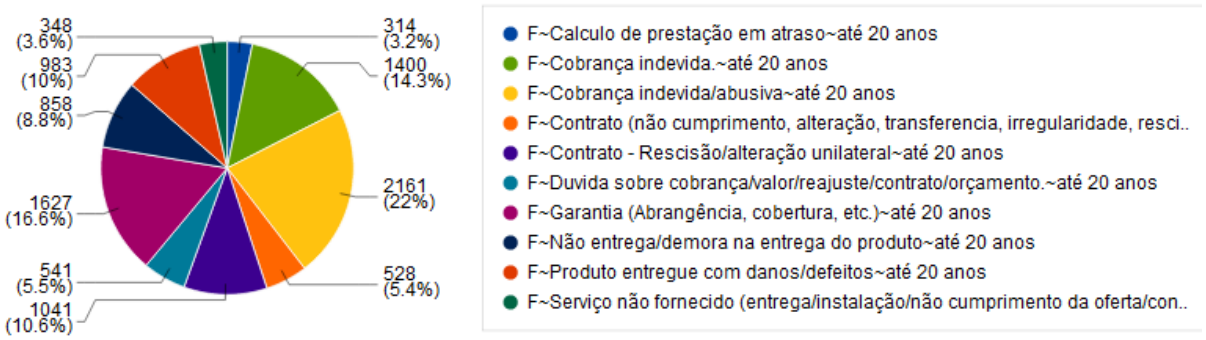
Figura 18- 10 principais problemas registrados por homens na faixa etária até 20 anos.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 19 – Mostra que existe uma semelhança em registro de reclamações, no tocante aos principais problemas, continua igual aos jovens de sexo masculino.

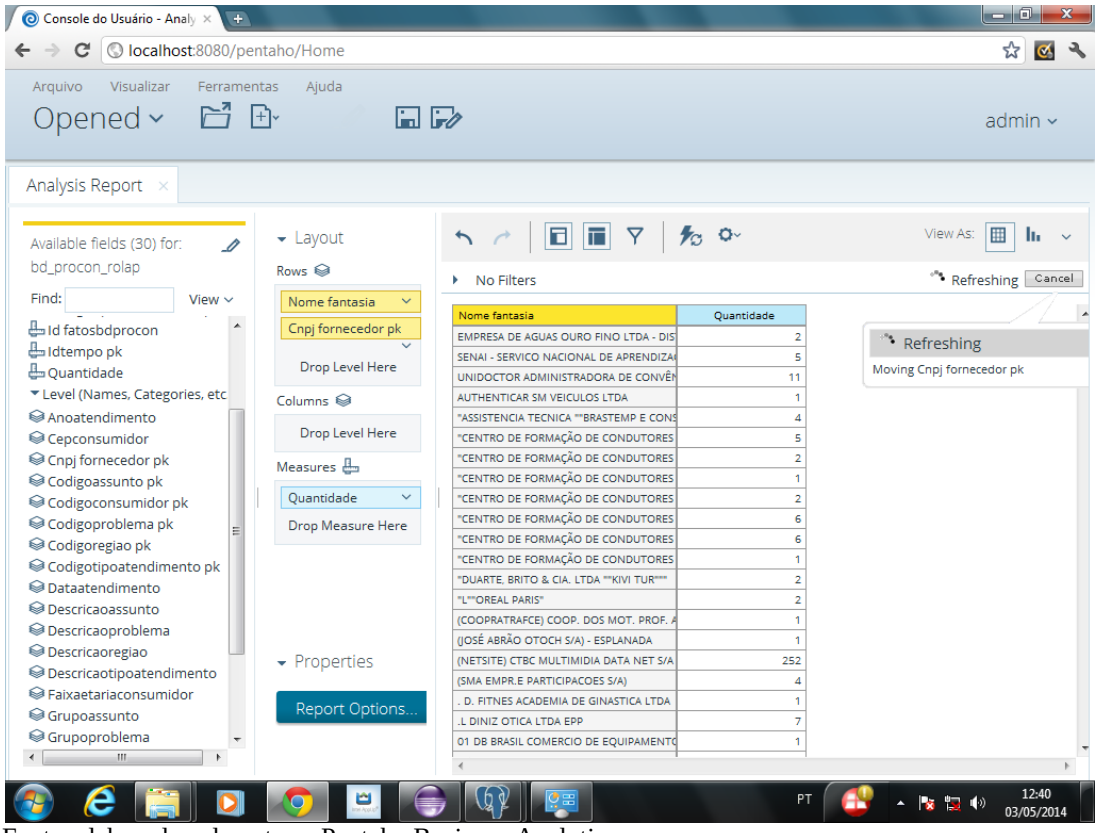
Figura 19 - 10 principais problemas registrados por mulheres na faixa etária até 20 anos.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

Na Figura 20 – Pode-se visualizar a construção de uma consulta utilizando-se o Pentaho Business Analytics.

Figura 20 - Pentaho Business Analytics. Criando Consultas.



Fonte: elaborada pelo autor – Pentaho Business Analytics.

6 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi extrair informações importantes dos dados disponibilizados pelo PROCON, no portal de Dados Abertos do Governo Federal. Tendo objetivo realizar análises nestes dados, utilizando algumas técnicas de extração de conhecimento, além de permitir a visualização dessas informações de maneira compreensível a qualquer cidadão, por meio de gráficos.

O objetivo de realizar as análises foi conseguido através, da criação de um Data Mart para conter os dados que seriam analisados, após terem passado pelo processo de Descoberta de Conhecimento em Bancos de Dados. As informações geradas foram através de gráficos, construídos com a ferramenta *Pentaho Business Analytics*, e para povoamento do banco de dados foi utilizado o *Pentaho Data Integration*. Quanto as análises realizadas a ferramenta se mostrou bastante eficiente, e intuitiva na forma de realizá-las, não necessitando a criação de consultas complexas utilizando SQL.

Quanto ao *Data Mart* criado no modelo estrela, tornou as consultas realizadas ao mesmo, mais fáceis e eficientes, melhorando assim o desempenho no processamento das mesmas, sendo então assim uma boa escolha o modelo utilizado.

No que diz respeito ao Sistema de Recomendação em que foi estudada a sua viabilidade. Os dados mostraram-se em não conformidade durante a criação da matriz de utilidade, que seria utilizada com o objetivo de possibilitar uma recomendação aos usuários da aplicação, sobre quais fornecedores tem perfil de consumidor insatisfeito que está mais distante ou mais próximo perfil de consumidor do usuário da aplicação. Pode-se verificar isto nos gráficos gerados, onde os fornecedores em geral têm um percentual no total de reclamações registradas muito semelhante. E isto acaba inviabilizando a recomendação, pois não há margens percentuais confiáveis para fazer uma recomendação precisa, e isto pode ser comprovado na matriz criada neste trabalho.

Finalmente, as análises feitas mostram-se relevantes, das informações contidas nos dados, cabendo então a verificação e realização dos trabalhos futuros a serem desenvolvidos, e disponibilização de mais informações.

Uma contribuição desse trabalho foi ainda o minicurso ministrado no Workshop de Tecnologia da Informação do Sertão Central (WTISC 2014), pois outros estudantes de Tecnologia da Informação poderão dar continuidade a este trabalho, ou realizar novos estudos

com dados disponíveis por outros órgãos, transmitindo assim conhecimento e informações legíveis a população.

7 TRABALHOS FUTUROS

Diante do trabalho de análise realizado, como trabalhos futuros serão realizadas novas análises considerando diferentes anos referentes aos registros das reclamações. Isso poderá ser realizado quando os dados dos outros anos forem disponibilizados no portal de dados abertos. Inclusive, tais dados já foram requeridos pelo autor. Essas análises por ano, serão facilmente realizadas visto que a proposta deste trabalho foi de um Data Mart.

Outro ponto a acrescentar, seria a automatização da divisão de arquivos para popular o *Data Mart* proposto.

Quanto ao sistema de recomendação, a verificação de todos os fornecedores, para construção de uma matriz de utilidade, que possa servir ao propósito de recomendar. Isto sendo feito de maneira automatizada.

No que diz respeito às análises, os dados poderão ser cruzados com dados do Produto Interno Bruto (PIB), e dados da população através do Instituto Brasileiro de Geografia e estatística (IBGE), gerando assim análises mais ricas sobre os dados do PROCON.

Por fim, verificar a possibilidade da criação de uma aplicação, para disponibilizar as informações obtidas com as analises, em aplicações web ou móvel, com o intuito de garantir que população e as empresas possam ter acesso a estas. Possibilitando interagir, de modo a realizar consultas sobre serviços de seus interesses. Podendo assim verificar a qualidade e avaliar os fornecedores de diversos serviços.

REFERÊNCIAS

APLICATIVOS e serviços que utilizam dados abertos. 2013. Disponivel em: <http://dados.gov.br/aplicativos/>. Acesso em: 10 set. 2013.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 03-13, 2011.

BICO, Fernanda C. et al. Legibilidade em Dados Abertos: uma experiência com os dados da Câmara Municipal de São Paulo. In: Simpósio Brasileiro de Sistemas de Informação, 8., 2012, São Paulo. **Anais...** São Paulo, 2012. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2012/0068.pdf>. Acesso em: 05 set. 2013.

CALDEIRA, Carlos Pampulim. Data Warehousing: Conceitos e Modelos. 2. ed.Silabo, 2012.

CAZELLA, Silvio César et al. Recomendação de Objetos de Aprendizagem Empregando Filtragem Colaborativa e Competências. In: XX SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 20., 2009, Florianópolis. **Anais do Simpósio Brasileiro de Informática na Educação.** Florianópolis: SBIE, 2009.

CODD, E. F.; CODD, S. B.; SALLEY, C. T. Providing OLAP to User-Analysts: An IT Mandate.E. F.Codd&Associates,1993. Disponível em <http://dev.hyperion.com/resource_library/white_papers/providing_olap_to_user_analysts.pdf >. Acesso em: 10 out. 2013.

DE ALMEIDA, Simone; MARÇAL, Rui Francisco Martins. Data mining na web para inteligência competitiva. In: **Simpósio de Engenharia de Produção**, 11., 2004, Bauru, SP, Brasil. **Anais...** São Paulo, 2004. Disponível em: <http://www.pg.utfpr.edu.br/ppgep/Ebook/ARTIGOS/2.pdf>. Acesso em: 29 ago. 2013.

DIGIAMPIETRI, Luciano A. et al. Minerando e caracterizando dados de currículos lattes. Proceedings of BraSNAM, 2012. In: **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**, 2012, Curitiba, PR, Brasil. Disponível em: <http://www.researchgate.net/publication/236212663_Minerando_e_caracterizando_dados_de_currculos_lattes/file/e0b49517024a3711a4.pdf>. Acesso em: 29 ago. 2013.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. Advances in Knowledge Discovery and Data Mining. Menlo Park, EUA: AAAI Press, 1996. 611 p.

FERREIRA, Jorge B., FACULDADE DEENGENHARIA ELÉTRICA, PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO .Mineração de Dados na Retenção de Clientes em TelefoniaCelular. 2005. 93 f. Dissertação (Mestrado em Engenharia Elétrica) -,

Faculdade deEngenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005.

JI, Changqing et al. BIG DATA PROCESSING: BIG CHALLENGES AND OPPORTUNITIES. **Journal Of Interconnection Networks.** {s.l.}, p. 1-19. dez. 2012. (JI. Et al., 2012)

KOTIDIS, Yannis; ROUSSOPOULOS, Nick. An alternative storage organization for ROLAP aggregate views based on cubetrees. In: **ACM Sigmod Record**. ACM, 1998. p. 249-258.

LICHTNOW, Daniel; GARIN, Ramiro Saldana; PALAZZO, Luis A. Moro. O uso de Técnicas de Recomendação em Um Sistema Para Apoio À Aprendizagem Colaborativa. **Revista Brasileira de Informática na Educação,** Porto Alegre - Rs, v. 14, n. 3, p.49-59, set. 2006.

LIMA, A. M. Adapitive virtual partitioning for OLAP query processing in a database cluster.In: Brasilian Symposium on DataBases (SBBD), 19., 2004, Brasilia,DF,Brasil**Anais...**Brasilia, 2004, Disponível em:<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.3757&rep=rep1&type=pdf >. Acesso em: 10 out 2013. 2004pp. 92-105.

MACHADO, Felipe Nery Rodrigues. **Tecnologia e Projeto de Data Warehouse: uma visão multidimensional**. Érica, 2006.

MACHADO, A. e PARENTE de Oliveira, J. Digo: An open data architecture for e government. In: In IEEE International Enterprise Distributed Object Computing Conference Workshops (EDOCW), 15., 2011, Helsink. **Anais....** Helsink: Dblp, 2011. p. 448 - 456.

MAGALHÃES, Cleyton Vanut Cordeiro de et al. Proposta de um Data Mart para avaliação de empresas usuárias do Twitter através das mensagens postadas pelos clientes. **Revista Brasileira de Administração Científica,** Aquidabã, v. 5, n. 3, p.123-135, ago. 2012.

OPEN KNOWLEDGE FOUNDATION OKFN. Open Data – An Introduction. [S.l.:S.n.], 2004. Disponível em: <http://okfn.org/opendata/>. Acesso em 09 out. 2013.

PEREIRA, Danilo Costa; ALEMIDA FILHO, Augusto Alves de; SOUZA, Ellen. Proposta de Data Mart para análise de reclamações realizadas em Procons. In: XIII JORNADA DE ENSINO, PESQUISA E EXTENSÃO – JEPEX 2013, 13., 2013, Recife. **Anais... Recife - PE.** Recife: Ufrpe.

PENTAHO: Any Data. Any Analytics. Simplified. 2014. Pentaho Corporation. Disponível em: <http://www.pentaho.com>. Acesso em: 05 jun. 2014.

RAJARAMAN, A; ULLMAN, J. D. **Mining of massive datasets.** ed. Cambridge: Cambridge University Press. 2012. 353p.

RAMAKRISHNAN, Raghu; GEHRKE, Johannes. **Database Management Systems**. 2. ed. Osborne/mcgraw-hill, 2000.

REZENDE, Solange Oliveira. **Sistemas Inteligentes:** Fundamentos e Aplicações. Barueri, SP: Manole Ltda, 2003. 525 p.

SILVA, Marcelino P. Santos. **Mineração de Dados-Conceitos, Aplicações e Experimentos com Weka**. In: Artigo. Instituto Nacional de Pesquisas Espaciais (INEP),2004. São José dos Campos-SP.

SILVA, Ticiana Linhares Coelho da. Análise em Big Data e um Estudo de Caso utilizando Ambientes de Computação em Nuvem. In: **Simpósio Brasileiro de Banco De Dados Sbbd**, 28., 2013, Recife: Sbbd, 2013.

SONG, I.- Y. Data Warehousing Systems: Foundations and Architectures. In: AND, L. L.; ÖZSU, M. T. (Eds.). **Encyclopedia of Database Systems** . Springer US, 2009. p. 684-692.

THE OLAP COUNCIL. OLAP and OLAP Server Definitions. 1995. Disponível em:

<http://dssresources.com/glossary/olaptrms.html>. Acesso em 17 out. 2005.

WU, Xindong et. al. Data Mining with Big Data. **Knowledge and Data Engineering, IEEE Transactions**, Boston, p. 1-25. 26 jun. 2013.