UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE GRADUAÇÃO DE TECNOLOGIA EM SISTEMAS PARA INTERNET

DANIEL DELAPORTE

PREPARAÇÃO DE DADOS DE PLANO DE SAÚDE SUPLEMENTAR PARA ALGORITMOS DE MINERAÇÃO DE DADOS

TRABALHO DE CONCLUSÃO DE CURSO

CAMPO MOURÃO - PR

2015

DANIEL DELAPORTE

PREPARAÇÃO DE DADOS DE PLANO DE SAÚDE SUPLEMENTAR PARA ALGORITMOS DE MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação de Tecnologia em Sistemas para Internet da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de Tecnólogo em Tecnologia em Sistemas para Internet.

Orientador: Prof. Me. Everton Fernando Barros

Co-orientador: Prof. Dr. Diego Bertolini Gonçalves

CAMPO MOURÃO - PR

2015

Dedico este trabalho de conclusão de curso aos meus pais, Milton Delaporte e Valdete Pereira Delaporte; ao orientador pela ajuda no decorrer do trabalho; aos professores e amigos.

AGRADECIMENTOS

Gostaria primeiramente de agradecer a Deus por ter me dado saúde e persistência para superar todos os obstáculos e sempre iluminar o meu caminho.

Aos meus pais Milton Delaporte e Valdete Pereira Delaporte, ao meu irmão André Delaporte, e toda minha família que sempre estiveram presentes, compreenderam as minhas ausências nos momentos de estudo e nunca mediram esforços para que eu pudesse alcançar este degrau da minha vida, pelo carinho, incentivo e amor.

Agradeço a todos os professores que me conduziram durante a graduação, em especial ao meu orientador Prof. Me. Everton Fernando Barros, por toda a paciência, dedicação e confiança em mim depositadas, fatores únicos esses que, possibilitaram a realização e conclusão deste trabalho.

E a todos que direta ou indiretamente fizeram parte da minha formação e conquista, os meus sinceros agradecimentos.

RESUMO

DELAPORTE, Daniel. PREPARAÇÃO DE DADOS DE PLANO DE SAÚDE SUPLEMENTAR PARA ALGORITMOS DE MINERAÇÃO DE DADOS. 31 f. Trabalho de Conclusão de Curso – Curso de Graduação de Tecnologia em Sistemas para Internet, Universidade Tecnológica Federal do Paraná. Campo Mourão - PR, 2015.

O grande acúmulo de dados gerados pelos sistemas de informações que se fazem indispensáveis em todas as áreas torna possível o processo de Descoberta de Conhecimento em Banco de Dados (DCBD), porém esses dados precisam ser ajustados de alguma forma que seja possível extrair algum conhecimento potencialmente útil para o domínio em que se aplica. Desta forma, para obter conhecimento de grandes bases de dados utiliza-se as metologias de DCBD. Neste trabalho foi utilizado a metodologia proposta por Fayyad et al. (1996) com esta metodologia foi realizado o pré-processamento de uma parte dos dados do *data warehouse* de um Plano de Saúde Suplementar (PSS), afim de incluir registros contidos em determinada tabela como atributo de uma nova tabela, agrupando os eventos que foram realizados pelos beneficiários desse PSS. Ao final aplicando toda da metologia proposta e passando por todas as etapas, chegou-se como resultado a criação de uma ferramenta para manipular todo esse processo tornando possível a conversão dos registros dos eventos em atributos de uma tabela.

Palavras-chave: Preparação de Dados, Mineração de Dados, Inteligência de Negócios, Descoberta de Conhecimento em Banco de Dados.

ABSTRACT

DELAPORTE, Daniel. . 31 f. Trabalho de Conclusão de Curso – Curso de Graduação de Tecnologia em Sistemas para Internet, Universidade Tecnológica Federal do Paraná. Campo Mourão - PR, 2015.

The large accumulation of data generated by information systems that are essential in all areas makes possible the process of Knowledge Discovery in Database (KDD), but these data need to be adjusted in some way it is possible to extract some knowledge potentially useful for the domain in which it applies. Thus, to obtain knowledge of large databases is used the metologias of KDD. In this work we used the methodology proposed by Fayyad-1996b with this methodology was carried out pre-processing part of the data warehouse of a Prepaid Health Plans in order to include records contained in a particular table as an attribute of a new table, grouping the events that have been made by beneficiaries of Prepaid Health Plans. At the end applying all of the proposed Methodology and going through all the steps, it was reached as a result the creation of a tool to manipulate the entire process making it possible to convert the events of records attributes of a table

Keywords: Data Preparation, Data Mining, Business Intelligence, Knowledge-Discovery in Databases.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

CID Classificação Internacional de Doenças

MD Mineração de Dados

DCBD Descoberta de Conhecimento em Banco de Dados

BD Banco de Dados

DW *Data Warehouse*

PSS Plano de Saúde Suplementar

ANS Agência Nacional de Saúde Suplementar

CFM Conselho Federal de Medicina

SQL *Structured Query Language*

CSV *Comma-separated values*

TGE Tabela Geral de Eventos

SUMÁRIO

1 INTRODUÇÃO

O armazenamento de dados nos sistemas computacionais é uma prática comum e essencial, devido à queda do custo de armazenamento dos dados, e a rápida automatização das empresas, que fez com que a quantidade de dados armazenados em bases de dados crescesse em alta velocidade. Esse grande acúmulo e volume de dados existem nas mais variadas áreas, tais como: financeira, comercial, científica, produção, manufatura e médica (REZENDE, 2003).

Com o grande volume de dados disponíveis, nota-se que pode existir conhecimento importante e potencialmente útil para o domínio no qual se aplica. No entanto, descobrir esse conhecimento nem sempre é uma tarefa fácil.

A análise de grandes volumes de dados e a transformação de dados em conhecimento serve como base para apoiar a tomada de decisões em diversos segmentos, como no caso das operadoras de planos de saúde suplementar que possuem uma grande quantidade de dados armazenados, como procedimentos realizados pelos beneficiários. Essas bases podem conter conhecimentos ocultos de alta qualidade que auxiliam na tomada de decisões na gestão de planos de saúde e em programas de prevenção de doenças.

Uma das técnicas utilizadas para descoberta de conhecimento oculto em banco de dados consiste na aplicação de algoritmos de Mineração de Dados (MD), que faz parte do processo de Descoberta de Conhecimento em Banco de Dados (DCBD). No entanto, não importa o quão bom o algoritmo de MD seja, ele falhará nessa descoberta se a qualidade dos dados for ruim. Como a maioria das bases de dados apresenta problemas, (valores inconsistentes, falta de precisão, ruídos e erros de medição) que reduzem sua qualidade, surge à necessidade de se aplicar técnicas de pré-processamento para melhoria de qualidade desses dados.

Em meio aos diversos problemas encontrados em bases de dados, realizou-se neste estudo a aplicação de técnicas de pré-processamento em dados de um Plano de Saúde Suplementar (PSS), com o objetivo de prepará-los para utilização em algoritmos de MD, esses dados foram obtidos exclusivamente para fins de pesquisa, portanto toda e qualquer informação pessoal foi retirada, contendo assim somente informações administrativas e de procedimentos em

hospitais, laboratórios e consultórios relativos aos beneficiários de um PSS do Estado de Santa Catarina (BARROS et al., 2011).

A técnica de DCBD proposta por Fayyad et al. (1996) consiste em nove etapas, porém devido ao objetivo do trabalho que é o pré-processamento, será fundamentado nas cinco primeiras etapas, que são: 1 - Entendimento do Domínio da Aplicação, 2 - Seleção de um Conjunto de Dados Alvo, 3 - Limpeza de Dados, 4 - Pré-seleção de Atributos e 5 -Redução e Projeção de dados.

No que se refere às operadoras de planos de saúde suplementar, estas possuem uma grande quantidade de informações armazenadas em seus bancos de dados, muitas vezes com informações sobre os procedimentos realizados com seus beneficiários. Esses dados podem revelar informações ocultas, que podem ser de grande utilidade em tomadas de decisões tanto para prevenção de doenças quanto para redução de custos com os tratamentos (BARROS et al., 2011).

O objetivo desse trabalho de conclusão de curso é desenvolver uma ferramenta que ajude a preparar uma parte do *data warehouse* (DW) cedido, mais especificadamente as tabelas `TGE` (Tabela Geral de Eventos) que contém a descrição dos procedimentos realizados pelos beneficiários, e `FATO_EVENTO` (Tabela Eventos Realizados) que contém a ocorrência dos procedimentos realizados pelos beneficiários, preparando apenas esse conjunto de dados para uma futura aplicação de algoritmos de MD, que pode resultar na descoberta de informações úteis para gestores de planos de saúde suplementar.

Este trabalho está estruturado da seguinte forma: no primeiro capítulo é apresentada a introdução; no segundo, o referencial teórico apresenta os conceitos relacionados com este trabalho; no terceiro, a identificação da necessidade sobre a preparação de dados para indução de CID; no quarto, os resultados obtidos com o pré-processamento; e por fim, no quinto capítulo, são apresentados as conclusões e trabalhos futuros.

2 REVISÃO BIBLIOGRAFICA

Este trabalho utilizou o DW cedido por um PSS de Santa Catarina, onde será aplicada a técnica de pré-processamento proposta por Fayyad et al. (1996), utilizando o algoritmo de Levenshtein, para auxiliar na busca de semelhança entre as descrições dos eventos realizados pelos beneficiários do PSS, Alves et al. (2013) utilizou a técnica de DCBD de Fayyad et al. (1996) combinada com algoritmo de Levenshtein entre outros, para encontrar alinhamento entre ontologias, Silva et al. (2011) utilizou o processo de DCBD de Fayyad et al. (1996) combinado com algoritmo de Levenshtein entre outros para detecção de registros duplicados em acervos de bibliotecas.

2.1 PLANOS DE SAÚDE SUPLEMENTAR

O plano de saúde suplementar (PSS) pode ser definido como todo atendimento privado de saúde, realizado ou não por meio de um convênio com um plano de saúde. Estão presentes dentro do cenário da Saúde Suplementar no Brasil o governo representado pelo Ministério da Saúde, a Agência Nacional de Saúde Suplementar (ANS) e a Agência Nacional de Vigilância Sanitária (ANVISA) - além das operadoras de planos privados, as seguradoras e os prestadores de serviço de assistência à saúde.

Com a criação da ANS (Agência Nacional de Saúde Suplementar) em novembro de 1999 pela medida provisória N.1926, convertida na lei N.9.961, como órgão de regulação, normatização, controle e fiscalização das atividades que garantem a assistência suplementar à saúde ocorreram mudanças na estrutura do setor de saúde suplementar, no qual agora o modelo de atenção é focado nas ações de promoção e prevenção de doenças e não mais centrado na doença, além de prever uma ampliação na utilização de sistemas de informação como insumo estratégico (BRASIL, 2000).

Com a utilização desses sistemas de informação o acumulo de dados é cada vez maior. Sendo a otimização dos processos operacionais cada vez mais um fator crítico para o sucesso das organizações que atuam na área de saúde, a tecnologia de informações ocupa cada vez mais

um papel de destaque entre as ferramentas de gerenciamento na gestão de saúde.

2.2 DCBD - MÉTODO DE FAYYAD

A metodologia proposta por Fayyad et al. (1996) explica que a Descoberta de Conhecimento em Banco de Dados (DCBD), consiste em um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis implícitos nos dados. O método de Fayyad et al. (1996) contém nove etapas, porém como o objetivo desse trabalho é apenas o pré-processamento, este trabalhado utilizara apenas as cinco primeiras etapas do processo de DCBD proposto por (FAYYAD et al., 1996), pois são essas as etapas responsáveis pelo pré-processamento..

1. Entendimento do domínio da aplicação.
2. Seleção de um conjunto de dados alvo.
3. Limpeza de dados.
4. Pré-Seleção de atributos.
5. Redução e Projeção de dados.
6. Escolha do Algoritmo de Mineração de Dados
7. Mineração de Dados
8. Interpretação dos padrões minados
9. Consolidação do conhecimento descoberto

A Figura 3 exemplifica as sequências das etapas do pré-processamento, e essas etapas serão explicadas nas seções a seguir.

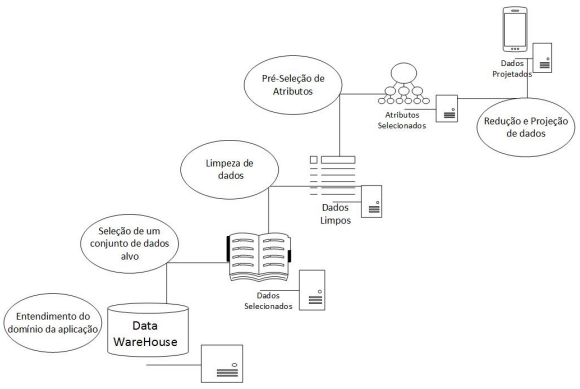

Figura 1: Fluxograma do pré-processamento

2.2.1 ENTENDIMENTO DO DOMÍNIO DA APLICAÇÃO

Na maioria dos casos o pré-processamento de dados em um processo de DCBD, exige um entendimento muito grande de conhecimento do domínio. É fundamental que antes de se iniciar o pré-processamento, que se tenha o conhecimento e entendimento dos dados, evitando assim a busca por dados irrelevantes e sem relação importante para o objetivo determinado, como por exemplo, a tabela `TGE` contém atributos como `CIRURGICO` e `ESTRUTURA` e a tabela `FATO_EVENTO` contém atributos como `ID_CLASSE_GERENCIAL_PF` que são atributos irrelevantes para esse trabalho.

2.2.2 SELEÇÃO DE UM CONJUNTO DE DADOS ALVO

Depois de realizado o entendimento do domínio, deve-se focar no conjunto de dados ou subconjunto de atributos que será realizada o DCBD. A Seleção de um Conjunto de Dados Alvos consiste de decidir sobre quais dados serão utilizados na MD, levando em consideração os atributos, e instâncias das tabelas de acordo com sua relevância para MD, sua qualidade e suas limitações técnicas, como por exemplo, volume e/ou tipo de dados.

2.2.3 LIMPEZA DE DADOS

Nem todos os dados armazenados em banco de dados estão totalmente coerentes, normalmente eles estão incompletos, inconsistentes ou apresentam algum outro tipo de ruído, o processo de limpeza de dados tem como objetivo eliminar estas anomalias dos dados, para aumentar e melhorar a qualidade destes.

2.2.4 PRÉ-SELEÇÃO DE ATRIBUTOS

Ao analisar um grande conjunto de dados pode haver centenas de atributos, alguns deles irrelevantes ou até mesmo redundantes para a MD.

Tomando por exemplo a análise de que um determinado beneficiário possa apresentar determinado tipo de doença e que será necessário realizar algum procedimento com o mesmo, o atributo `SITUACAO_RH` será irrelevante para essa classificação, em contrapartida os atributos `DATA_NASCIMENTO` e `SEXO`, são potencialmente relevantes para essa classificação, pois se sabe que algumas doenças tende a se desenvolver dependendo da idade e sexo do individuo.

A pré-seleção de atributos busca também remover a dimensionalidade dos dados, isto é a quantidade de atributos utilizados para descrever um conjunto de instâncias. Há várias

razões que podem justificar a redução do número de atributos no caso de excessividade de dimensionalidade, dentre elas, acurácia, tempo e espaço, relevância, redundância e simplicidade (HAN et al., 2006);(RIAÑO, 1997).

- *Acurácia*: pré-processar o conjunto de atributos pode melhorar a acurácia do modelo resultante.
- *Tempo e Espaço*: o custo de tempo e espaço dos algoritmos de indução estão diretamente relacionado à quantidade de atributos.
- *Relevância*: alguns atributos considerados inúteis podem ser descartados.
- *Redundância*: atributos que contém a mesma informação representada de forma diferente em outro atributo também podem ser descartados.
- *Simplicidade*: a redução de atributos reflete significativamente na criação de estruturas menores, permitindo um melhor entendimento do domínio.

2.2.5 REDUÇÃO E PROJEÇÃO DE DADOS

Ao fazer a redução e projeção dos dados procura-se encontrar características úteis para fazer a representação dos dados conforme seja o objetivo da tarefa, objetivando a redução do número de variáveis e ou instâncias a serem consideradas no conjunto de dados, preservando a integridade original dos dados bem como o enriquecimento semântico das informações. Essa redução do conjunto de dados pode deixar a MD mais eficiente e ainda capaz de produzir os mesmo resultados (ou quase os mesmos), (HAN et al., 2006) (FAYYAD et al., 1996). Há várias técnicas e estratégias de redução de dados. Algumas técnicas são apresentadas conforme (BARROS et al., 2011):

- Agregação em cubos de dados, em que operações de agregação são aplicadas sobre os registro para construção de um cubo de dados.
- Seleção de um subconjunto de atributos, no qual atributos irrelevantes, fracamente relevantes ou redundantes devem ser detectados e removidos.
- Redução de dimensionalidade, em que mecanismos de codificação para reduzir o tamanho de dados.
- Redução de numerosidade, no qual os dados são substituídos ou estimados por alternativas, representação de dados menores tais como modelos paramétricos (que armazenam

apenas os parâmetros do modelo em vez dos dados reais) ou métodos não paramétricos tais como agrupamento, amostragem e o uso de histogramas.

- Discretização e geração de conceitos hierárquicos, no qual dados brutos dos atributos são substituídos por níveis conceituais mais altos. Discretização dos dados é a uma maneira de redução de numerosidade que é muito útil para geração automática de conceitos hierárquicos. Discretização e geração de conceito hierárquico são poderosas ferramentas para mineração de dados, na medida em que permitem mineração de dados em múltiplos níveis de abstração.

2.3 ALGORITMO DE LEVENSHTEIN PARA DISTÂNCIA

O algoritmo Levenshtein *Distance* foi criado pelo russo Vladimir Levenshtein, em 1965, o foco da sua técnica é a avaliação de similaridade entre duas *strings* com base no número de operações que precisam ser realizadas para se transformar uma *string* em outra, utilizando operações de inserção, exclusão e substituição. Quando a distância entre as *strings* for igual a zero, elas são idênticas. Dadas as *strings* a serem comparadas é montada uma matriz, onde é informado os custos de cada operação. Ao final das comparações se obtêm a distância pela última posição da matriz (KOUYLEKOV; MAGNINI, 2005).

O algoritmo de Levenshtein é parafraseado como “o menor número de inserções, remoções e substituições para igualar duas *strings*” (GONDIM F, 2006).

Essa técnica foi utilizada por se tratar de uma técnica que analisa as *strings* comparando-as carácter a carácter, sendo assim quando a distancia entre a comparação de duas *strings* é igual a zero, teremos certeza de que se trata de uma igualdade de fato.

A Figura 2 mostra um exemplo da matriz de distâncias de duas palavras SITTING e KITTEN

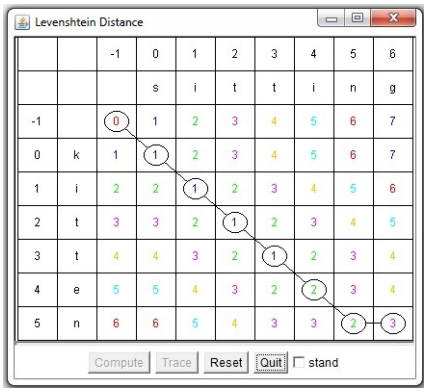

Figura 2: Algoritmo de Distância de Levenshtein

O algoritmo de Levenshtein indica a distância entre as duas *strings* com um valor de 3, essa é a quantidade de edições necessárias para que ele consiga transformar uma *string* em outra. Ao comparar o caractere K da palavra *Kitten* com ocaractere S da palavra *Sitting*, há a necessidade de fazer a edição do caractere para transformar um caractere no outro, e a distância entre as duas palavras ganha o valor de 1, a próxima comparação é entre o caractere I da palavra *Kitten* e o caractere I da palavra *Sitting*, como se trata do mesmo caractere não há a necessidade de edição e a distância entre as duas palavras ainda continua com o valor de 1, até o momento em que se compara o caractere E da palavra *Kitten* com o caractere I da palavra *Sitting*, há novamente a necessidade de fazer uma edição para transformar um caractere no outro, então a distância entre as duas palavras é incrementada em mais uma edição passando a valer 2, a próxima comparação é o caractere N com o caractere G que necessita novamente de mais uma edição para transformar um carácter no outro, sendo assim a distância é incrementada em mais uma edição passando a valer 3, tendo como resultado o valor de 3 para a distância entre as palavras *Kitten* e *Sitting*

O próximo capitulo tratará do desenvolvimento do trabalho utilizando o metodologia apresentada.

3 PRÉ-PROCESSAMENTO

Neste capítulo, é apresentado o desenvolvimento do trabalho aplicando a metodologia apresentada por Fayyad et al. (1996) juntamente com algoritmo de Levenshtein *Distance* no DW cedido pelo PSS de Santa Catarina. O processo de DCBD é interativo e iterativo e envolve repetições de etapas com muitas decisões que devem ser tomadas no decorrer do processo, portando, essas etapas podem ser refeitas ou refinadas conforme o desenvolvimento do trabalho.

Em estudo realizado por Barros et al. (2011) alguns resultados na preparação desses dados foram obtidos entretanto, na época não existia um DW desses dados da forma que foi recebido, com os dados mais limpos e organizados acredita-se que melhorará a preparação desses dados, além de possuir um objetivo diferente da preparação realizada anteriormente.

A Tabela 1 apresenta a principal tabela do DW em que possui a ocorrência dos eventos realizados pelos beneficiários e a Tabela 2 apresenta como deverá ser a tabela final após o pré-processamento.

ID_EVENTO	ID_TGE	ID_BENEFICIARIO	ID_CID	ID_DATA_EVENTO
27	4647	31511	13027	18586
28	4730	31511	13027	18586
36	2357	31511	13027	18586
37	2370	31511	13027	18586
38	3490	6766	13027	18587
39	4687	4679	13027	18573
40	4700	4679	13027	18573
41	3490	4679	13027	18573
45	3588	4679	13027	18573
46	4557	19794	13027	18573

Tabela 1: Tabela `FATO_EVENTO` **do** `DWSAUDE`

Entre os vários atributos que se encontra na tabela `FATO_EVENTO` do DW, os atributos que tem relevância para este trabalho são os seguintes: `ID_EVENTO` é a identificação da ocorrência do evento, `ID_TGE` que é a identificação do evento realizado, `ID_BENEFICIARIO` é a identificação de cada beneficiário, `ID_CID` que é a identificação para a CID constatada, `ID_DATA_EVENTO` é a identificação para a data em que ocorreu o evento.

Em trabalhos anteriores Barros et al. (2011) realizou-se MD com os dados dispostos como a Tabela 1, entretanto os resultados não foram satisfatórios, viu-se então no campo ID_TGE a oportunidade de se obter dados melhores para MD, uma vez que acredita-se que o campo ID_TGE possui alto poder preditivo, mas não da forma como esta organizado.

Portanto esse trabalho busca transformar os registros que existem na coluna ID_TGE em atributos de uma nova tabela, passando primeiro por uma fase de eliminação de redundância dos eventos existentes, em seguida passa pela fase de agrupamento de eventos criando grupo para determinados tipos de eventos em seguida cada atributo de evento geral será relacionado ao seu grupo, conforme atendimento realizado por dia e beneficiário, ao final do trabalho a ferramenta permitirá a geração da Tabela 2, para isso será aplicada a metodologia apresentada anteriormente. O significado das siglas dos atributos da Tabela 2 pode ser encontrado no apêndice A.

ID_BENEF	ID_DATA	GE1	GE2	GE3	GEN	ID_CID
31511	18586	3478	4746	4730	4647	13027
38555	18956	2357	3490	4730	3474	145
21345	18734	4851	4687	4730	4700	12467
27135	18345	3478	4746	4730	3588	85
28457	19345	2357	3490	4730	3474	1745
30908	15385	3478	2348	4730	3474	19256

Tabela 2: Tabela FATO_EVENTO_GRUPO

3.1 ENTENDIMENTO DO DOMÍNIO DA APLICAÇÃO

O DW cedido é oriundo de um PSS do estado de Santa Catarina, nele estão contidas algumas informações sobre os seus beneficiários, como tipo de contrato, o regime de atendimento, finalidade do atendimento, eventos realizados dentre outros. Observando todas essas informações surgiu à necessidade de desenvolver uma ferramenta para auxiliar no pré-processamento dos dados, dispondo-os de forma agrupada para posteriormente utilizar algoritmos de MD a fim de encontrar padrões de CID. CID significa Classificação Internacional de Doenças, ele fornece códigos relativos à classificação de doenças de uma grande variedade de sinais, sintomas, aspectos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças. Para cada tipo de doença ou sintoma o médico pode atribuir ou não um código de CID para o beneficiário de plano de saúde. Conforme esclarece o Código de Ética Médica:

O art. 73 do Código de Ética Médica, veda ao médico: "Revelar o fato de que tenha conhecimento em virtude do exercício de sua profissão, salvo por justo dever ou consentimento, por escrito, do paciente", 1- Ética médica – código. I. Título. II - Resolução CFM nº 1931, de 17 de setembro de 2009 (MÉDICA, 2009).

As CID, presentes no DW, parte do principio que o paciente autorizou o médico a informar, e será utilizada junto com outros atributos que possibilitem a indução de CID. Vale ressaltar que o DW não contem qualquer informações de cunho particular ou que permitam identificar um beneficiário do PSS.

3.2 SELEÇÃO DE UM CONJUNTO DE DADOS ALVO

Uma vez que utilizado um banco de dados relacional, o conjunto de dados alvo selecionado para realizar a descoberta de um CID, foram duas tabelas `TGE` e `FATO_EVENTO`, conforme eventos realizados com os beneficiários acredita-se que a informação útil para se encontrar esses padrões estejam no atributo `DESCRICAO` da tabela `TGE` e nos atributos `ID_TGE`, `ID_BENEFICIARIO`, `ID_DATA_EVENTO` e `ID_CID` da tabela `FATO_EVENTO`.

De posse do DW, foi iniciado o processo de criação do ambiente para manipular os dados, instalação do Sistema de Gerenciamento de Banco de dados (SGBD) *Sql Server 2008*, restauração do DW chamado `DWSAUDE` e instalação do Pentaho Data Integration, essa ferramenta é muito utilizada no ambiente de extração de dados e MD, pois se trata de uma ferramenta *open source* com vários recursos já prontos que agilizam a extração e manipulação de dados contidos em arquivos textos, banco de dados, planilhas dentro outros.

A Tabela geral de eventos `TGE` contém todos os eventos cadastrados manualmente pelos diversos usuários dos sistemas que alimentaram o banco de dados, que originaram esse DW, onde se encontra o foco desse trabalho que é o tratamento desses eventos podendo assim diminuir a quantidade de registros, eliminando duplicidades, entre outros problemas que podem conter essas informações. A tabela `FATO_EVENTO` contém todos os eventos que foram realizados com os beneficiários dos planos de saúde suplementar, essa tabela guarda entre outras informações, como, o `ID_TGE` que é o código da descrição do evento realizado com os beneficiários e `ID_CID` que é o código da Classificação Internacional de Doenças que o médico informou para aquele atendimento.

3.3 LIMPEZA DE DADOS

Nessa etapa o objetivo é eliminar as anomalias e redundâncias encontradas nas descrições da tabela `TGE`, gerando assim uma menor quantidade de registros. Essa redução é fundamental para o algoritmo que vai prever a CID, pois esses registros serão transformados em atributos, portanto a redução do número de registros implicará futuramente na redução no número de atributos.

Como dito anteriormente a tabela TGE contém a descrição dos eventos que podem ser feitos com os pacientes, descrição essa que não segue uma regra de nomenclatura, ou seja, não houve uma padronização no inicio do sistema que gerencia esses dados, com isso, é possível ter gerado muitos eventos que tem grande semelhança e ou até mesmo repetido, é nessas descrições que se pretende trabalhar ela é composta por 137.502 registros de eventos que podem ser realizados com beneficiários, já a tabela FATO_EVENTO é composta por 1.706.594 registros de eventos realizados com beneficiários, destes apenas 10014 eventos distintos da tabela TGE foram encontrados.

Baseado nessas informações identificou-se a necessidade de reduzir ainda mais esses números, pois ainda contém muitos eventos possivelmente duplicados e ou com ruídos. Para isso procuramos na literatura alguma forma de fazer essa análise que poderia resultar na unificação de algumas descrições de evento. Encontrou-se o algoritmo de Levenshtein, que durante a comparação entre duas palavras ou frases inteiras resulta em uma distância entre as duas, o que pode acusar uma semelhança ou proximidade entre elas exemplo: ID_TGE 28 Descrição SIFILIS (FTA-ABS) IGG e ID_TGE 19887 Descrição SIFILIS-FTA-ABS-IGG, são eventos que gerou uma distância de 3 entre eles, e o que muda de um para o outro é apenas os parênteses, ou seja, se trata do mesmo evento.

Existe um relacionamento entre a tabela FATO_EVENTO e a tabela TGE, que é o campo ID_TGE e utilizando a ferramenta Pentaho Data Integration, foi feita a extração dos 10014 eventos distintos da tabela TGE encontrados na tabela FATO_EVENTO, que originou o arquivo TGE_ID_DESCRICAO.CSV, esse arquivo é composto das seguintes informações: ID_TGE Identificador e Descrição, conforme Tabela 3.

ID_TGE	DESCRICAO
2	LEISHIMANIOSE,IFI PARA
3	LEPTOSPIROSE, AGLUTINACAO (MACRO E MICROSCOPIA)
6	LINFOCITOS T HELPER, CONTAGEM DE (IFI COM OKT-4) (CD4) CITOMETRIADE
7	LINFOCITOS T SUPRESSORES, CONTAGEM (IF COM OKT-8) (CD8) CITOMETRIADE
8	LISTERIOSE - AGLUTINACAO, POR ANTIGENO
13	MONONUCLEOSE – MONOTESTE
19	PROTEINA C REATIVA – DETERMINACAO QUANTITATIVA (TURBIDIME
20	PROTEINA C REATIVA, PESQUISA
22	RUBEOLA – HA
23	RUBEOLA – ANTICORPOS IGM, ELISA

Tabela 3: Arquivo TGE_ID_DESCRICAO.CSV

A partir desse passo, houve a necessidade da criação de uma ferramenta para ler esse arquivo e fazer a comparação dessas descrições utilizando o algoritmo de Levenshtein. Antes de enviar os eventos para o Algoritmo de Levenshtein e fazer a comparação foi feito o tratamento para a retirada de espaços entre os caracteres dos eventos, logo em seguida foi enviado

para o algoritmo de Levenshtein que retornava a distância entre os eventos que resultou em um novo arquivo chamado ComparacaoTratadaID.csv, conforme representado na Tabela 4. Esse arquivo é composto das seguintes informações: `ID1` Identificador Evento1, `Descricao1` Descrição Evento 1, `ID2` Identificador Evento 2, `Descricao2` Descrição Evento 2, `Distancia` Distância entre os eventos. O significado das siglas dos registros da Tabela 4 pode ser encontrado no apêndice B.

ID1	Descrição1	ID2	Descrição2	Distância
2	LSN	3	LEPT	33
2	LSN	6	LINFH	52
2	LSN	7	LINFS	54
2	LSN	8	LIST	26
2	LSN	13	MONON	20
2	LSN	19	PROT	58
2	LSN	20	PROTP	20

Tabela 4: Arquivo ComparacaoTratadaID.CSV

Como o universo de registros ainda é muito extenso, e que posteriormente esses registros se tornarão atributos, tentamos aproximar mais os eventos que poderiam ser provavelmente semelhantes, fechando mais o escopo para uma possível analise manual posterior, utilizamos a regra abaixo:

distancia <= menorDescricaoComparada / 2

Ou seja, distância entre os eventos comparados, deverá ser menor ou igual ao tamanho da metade da menor descrição de evento que está sendo comparado, sendo assim as descrições que não estiver dentro desse filtro presume-se que não há semelhança, foi utilizada a divisão por dois, pois é a primeira divisão que se pode fazer perdendo o mínimo de informação relevante, em experimento realizados com divisão por quatro e cinco, algumas descrições que continham abreviações como AG para Agulhas, ABD para ABDÔMEN, já causaria uma perca relevante. Esse processo resultou em mais um arquivo chamado MenorSTRdivididaID2.csv, conforme representado na Tabela 5, isso evita que descrições muito diferentes entre no escopo das analises manuais. O arquivo é composto das seguintes informações: `ID1` Identificador Evento1, `Descricao1` Descrição Evento 1, `ID2` Identificador Evento 2, `Descricao2` Descrição Evento 2, `Distancia` Distância entre Eventos. O significado das siglas dos registros da Tabela 5 pode ser encontrado no apêndice C.

ID1	Descrição1	ID2	Descrição2	Distância
6	LINFH	7	LINFS	14
20	PROT	3585	PROTJONES	12
23	RUBESIGM	24	RUBESIGG	1
23	RUBESIGM	19883	RUBEASIGG	8
23	RUBESIGM	19884	RUBEASIGM	7
24	RUBESIGG	19883	RUBEASIGG	7
24	RUBESIGG	19884	RUBEASIGM	8
28	SIFIGG	4794	SIFABS	4
28	SIFIGG	19887	SIFIGG	3
28	SIFIGG	19888	SIFIGM	4

Tabela 5: Arquivo MenorSTRdividdaID2.CSV

Em seguida foi criado um processo no Pentaho Data Integration (Figura 3) foi necessário criar esse processo para poder incluir os campos `ativo`, e `juntadoCom` para controle, que serão utilizados posteriormente, o processo inclui para cada linha do arquivo mais dois valores 1 e 0 separados por ponto e virgula respectivamente.

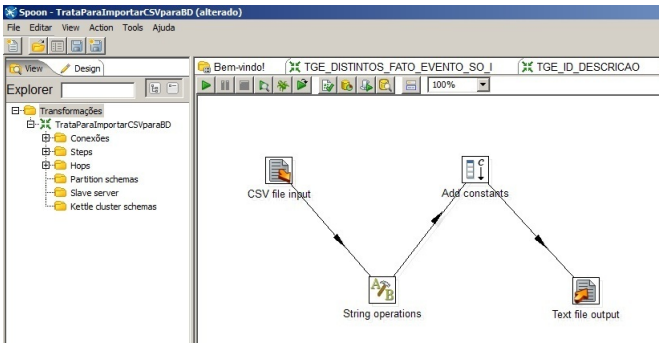

Figura 3: Processo de Tratamento das Descrições

Em seguida foi criado um novo processo no Pentaho Data Integration (Figura 4) para fazer a importação dessas informações contidas no arquivo MenorSTRdividdaID2.csv Tabela 5, resultando na criação de um banco de dados no MySQL, surgindo assim a base `BDSAUDE` e a tabela `DESCRICOES`. Essa tabela é composta pelas seguintes informações: Identificador, Identificado Evento 1, Identificado Evento 2, Descrição Evento 1, Descrição Evento 2, Distancia entre Eventos, Ativo e Juntado Com.

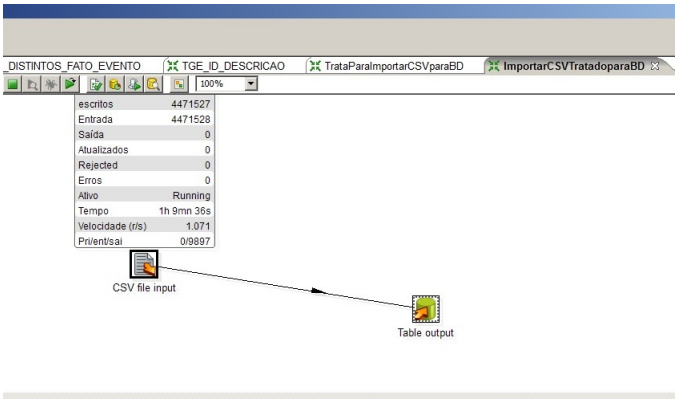

Figura 4: Processo de Importação das Descrições

Apenas o algoritmo de Levenshtein não é capaz de garantir a semelhança dos eventos, para poder unificar os mesmos, surge a necessidade de criar uma ferramenta para analisar caso a caso cada descrição, porém já filtrada pelos processos executados anteriormente.

Foi desenvolvido neste trabalho a aplicação ComparaDescricao.jar, que disponibiliza uma ferramenta que permite eliminar a redundância dos eventos que ainda faz parte do processo de redução de numerosidade dos eventos (Figura 5), seu objetivo foi possibilitar uma análise humano técnica especializada nas descrições dos eventos, quando identificado semelhança entre as descrições a ferramenta permitir marcar para unificar um evento ao outro solicitando um motivo, pois se trata de algo muito especifico da área da saúde. Para isso houve uma alteração na tabela `DESCRICOES` do banco de dados `BDSAUDE`, que foi a inclusão do campo motivo, alterando assim a composição da tabela `DESCRICOES` para: Identificador, Identificado Evento 1, Identificado Evento 2, Descrição Evento 1, Descrição Evento 2, Distancia entre Eventos, JuntadoCom e Motivo.

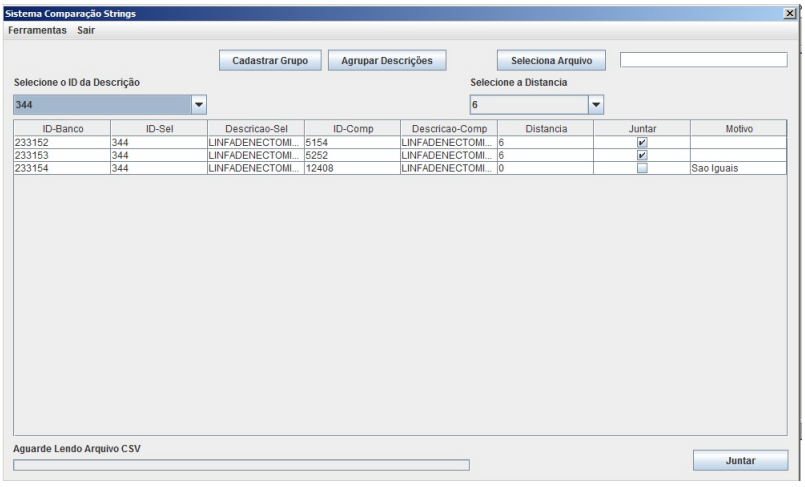

Figura 5: Tela de Eliminação de Redundância de Eventos

Ao finalizar essa etapa de análise manual para eliminar a numerosidade dos registros, deve-se gerar o arquivo de *update* para a tabela FATO_EVENTO, que fica disponível no menu Ferramentas, opção Gerar SQL *Update* FATO_EVENTO - DWSAUDE, e executá-lo no DWSAUDE, para substituir os registros da tabela FATO_EVENTO conforme Figura 6.

Figura 6: Menu para Gerar SQL *Update* FATO_EVENTO - DWSAUDE

Essa ferramenta vai disponibilizar o *script* SQL conforme Tabela 6.

/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 28 WHERE ID_TGE = 19887;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 34 WHERE ID_TGE = 19900;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 36 WHERE ID_TGE = 19902;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 48 WHERE ID_TGE = 21089;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 70 WHERE ID_TGE = 71;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 102 WHERE ID_TGE = 19836;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 114 WHERE ID_TGE = 21092;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 119 WHERE ID_TGE = 21113;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 152 WHERE ID_TGE = 12496;
/* Motivo: Sao Iguais */ UPDATE FATO_EVENTO SET ID_TGE = 161 WHERE ID_TGE = 12502;

Tabela 6: *Script* de *Update*_FATO_EVENTO.sql

3.4 PRÉ SELEÇÃO DE DADOS.

Essa etapa já faz parte da eliminação de dimensionalidade, onde os eventos vão passar a fazer parte de grupos, pois os eventos que estavam com problemas de redundância já foram eliminados no passo anterior, devemos agora voltar ao passo 3.3, fazer uma nova extração dos eventos distintos da tabela FATO_EVENTO, passar pelo algoritmo de Levensthein até surgir um novo arquivo ComparacaoTratadaID.csv.

Houve a necessidade de alterar novamente o banco de dados BDSAUDE, criando a tabela GRUPO, composta por : Identificador do Grupo, e Descrição do Grupo, e criado a tabela DESCRICOESGERAL, ficando assim composta por: Identificador, Identificado Evento 1, Identificador Evento 2, Descrição Evento 1, Descrição Evento 2, Distancia entre Eventos, Juntado Com, Motivo, Identificador de Grupo.

Utilizando o processo do Pentaho criado anteriormente para tratar e importar o arquivo csv, passamos o arquivo ComparacaoTratadaID.csv Tabela 4, para criação dos campos de con-

trole, e em seguida utilizamos outro processo do Pentaho (Figura 7), para importar todas as descrições para tabela `DESCRICOESGERAL`.

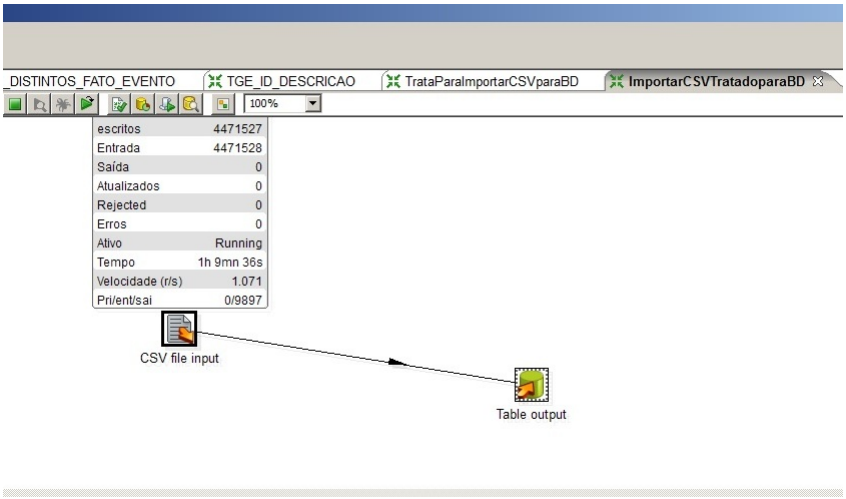

Figura 7: Processo de Importação do Arquivo CSV ComparacaoTratada.csv

Após a criação da ferramenta para eliminar a redundância de eventos, surgiu à necessidade de criar mais uma ferramenta dentro dessa mesma aplicação. Foi criada a ferramenta para diminuir a dimensionalidade dos eventos conforme a Figura 8, com ela podemos agrupar esses eventos a um domínio, possibilitando assim um melhor aproveitamento do algoritmo de indução de CID, pois a dimensionalidade será reduzida, isso impacta diretamente na quantidade de atributos que irão compor a tabela final.

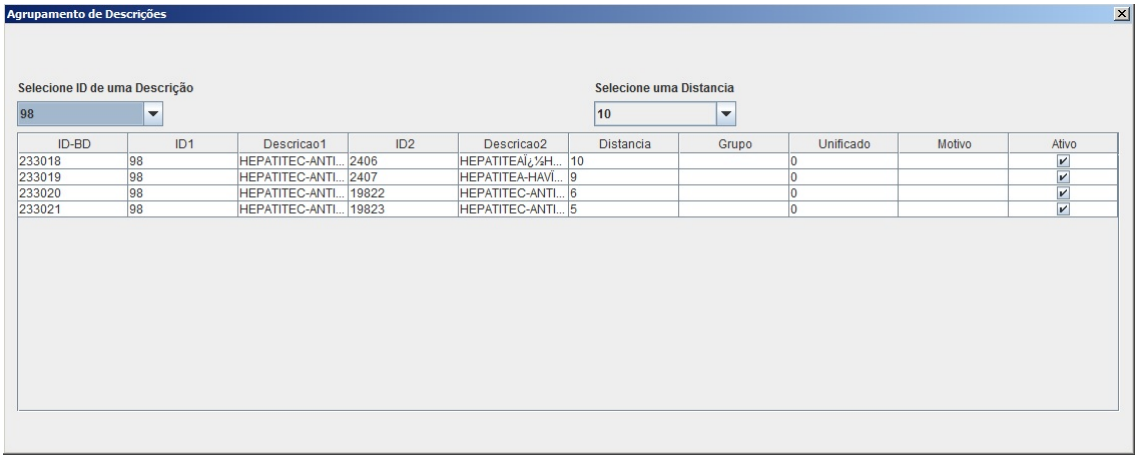

Figura 8: Tela de Eliminação de Dimensionalidade de Eventos

Para facilitar na análise manual, a tela da Figura 8 disponibiliza todas as descrições dos eventos comparadas com as outras descrições e da à distância entre elas, indicando assim uma possível semelhança, para serem agrupadas.

3.5 REDUÇÃO E PROJEÇÃO DE DADOS.

Ao concluir as etapas anteriores, devemos atualizar o `DWSAUDE`. Para isso foi criado dentro da aplicação ComparaDescricao.jar, alguns menus conforme Figura 9 que fazem a geração do `script` SQL, para efetivar as alterações no `DWSAUDE`. O apêndice D descreve a função de cada opção do menu da Figura 9.

Figura 9: Tela do Menu de Ferramentas

4 RESULTADOS

Esse trabalho resultou na criação de uma ferramenta que permite realizar a preparação de dados de um PSS, com foco na previsão de CID. Esta ferramenta permite fazer comparações entre textos, que nesse caso, foi utilizado no âmbito de pré-processamento de dados, para eliminar ruídos e inconsistências encontrados no `DWSAUDE` mais especificadamente na tabela `TGE`.

A ferramenta permite uma análise humano-técnica dos eventos ocorridos de forma mais sintetizada e objetiva, pois houve alguns filtros nas etapas do processo antes de chegar a análise manual.

Ao fim do processo, a ferramenta disponibiliza *scripts* SQL para efetivar as alterações necessárias no `DWSAUDE`, dentre eles um *script* que gerou um agrupamento dos eventos ocorridos conforme o atendimento realizado por dia e beneficiários.

Acredita-se que a forma como está disposta a nova tabela criada `FATO_EVENTO_GRUPO`, os algoritmos de indução de CID, podem ser mais efetivos.

O fluxograma da Figura 10 demonstra como deve ser utilizada a ferramenta.

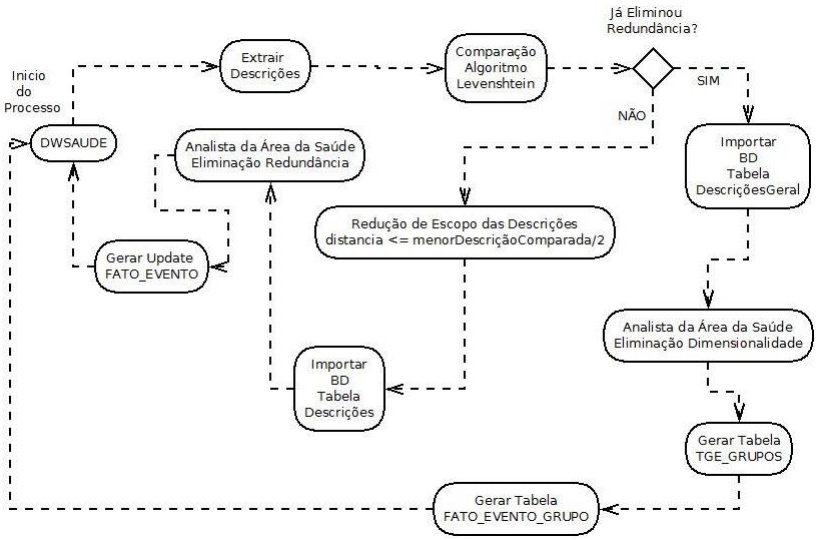

Figura 10: Fluxograma do processo realizado

O pré-processamento de dados demanda muitas vezes interação contínua e repetição dos processos. Iniciam-se os processos fazendo uma extração das descrições distintas existentes na tabela `FATO_EVENTO`, e passa essas descrições para o algoritmo de Levensthein calcular a distância entre elas, com o resultado dessas comparações, foi feita uma redução de escopo para aproximar mais as descrições semelhantes onde consideramos apenas as descrições que se atende a condição, distância <= menorDescriçãoComparada/2, o arquivo resultante desse filtro importamos para a tabela `Descricoes` do `BDSAUDE` criado localmente no MySQL.

No gráfico da Figura 11 é possível observar a quantidade de descrições que possuem total semelhança, com distância igual a zero, e alguma semelhança, com distância entre um e seis, e que são passíveis de análise pelos especialistas da área de saúde, podendo assim, reduzir consideravelmente o trabalho desses profissionais, pois reduz o escopo de descrições que devem ser analisadas.

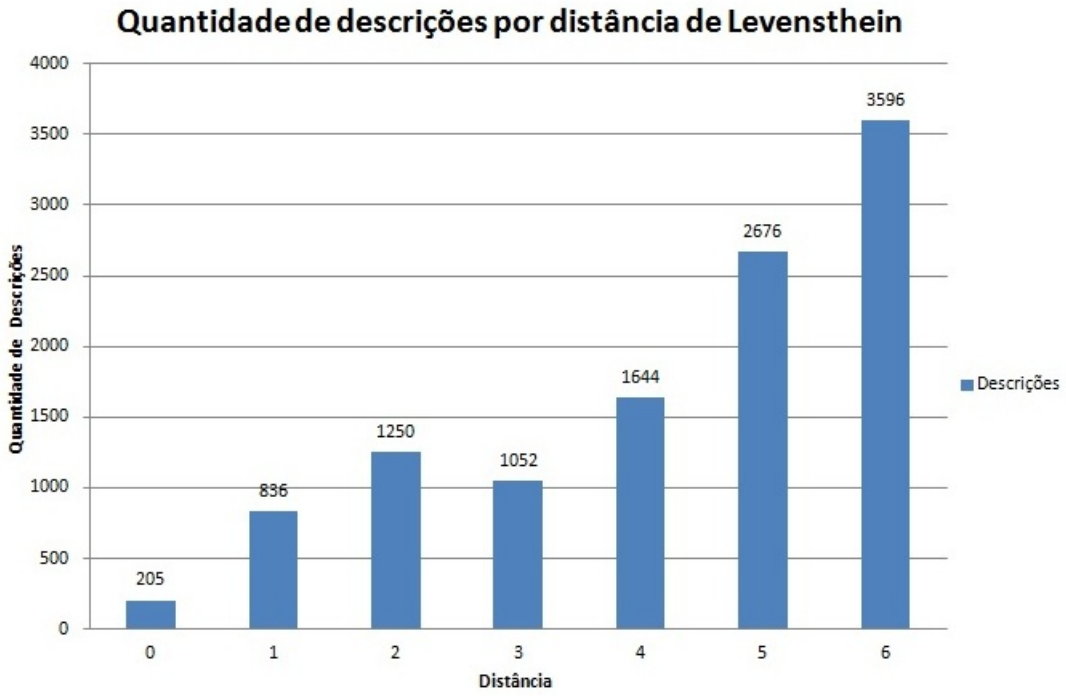

Figura 11: Quantidade de descrições por distância de Levensthein

Utilizando a ferramenta desenvolvida, o profissional da área da saúde pode fazer a eliminação de redundância entre as descrições e cada descrição que ele marcar para juntar, ele deve inserir um motivo (Figura 4). Efetuado a eliminação das redundâncias, fazemos a *update* da tabela `FATO_EVENTO` no `DWSAUDE`, com o *script* SQL resultante dessa operação.

Inicia-se o processo novamente, fazendo a extração dos dados, enviando para algoritmo de Levensthein calcular as distâncias, e como já se fez a eliminação de redundância, passamos

para a fase de eliminação de dimensionalidade dos atributos, importamos agora todo o arquivo resultante do algoritmo de Levensthein, para a tabela `DescricoesGeral`, para reduzir a dimensionalidade dos eventos, foi desenvolvido dentro da mesma ferramenta a tela para eliminar dimensionalidade (Figura 5). A ferramenta permite criar grupos e associar as descrições a eles, sendo assim diminuindo a dimensionalidade de atributos para a tabela final.

Após terminar o processo, atráves do menu de ferramentas da aplicação (Figura 9) é possível fazer a geração do *script* SQL de criação e inserção da tabela `TGE_GRUPOS`, que deve ser incluída no `DWSAUDE`, e em seguida se faz a geração do *script* SQL de criação e inserção da tabela `FATO_EVENTO_GRUPO`.

A Tabela 7 apresenta o resultado exemplificativo desse processo indicando assim que é possível gerar os dados necessários para utilização em um algoritmo de mineração de dados. De acordo com um dos objetivos do trabalho. O significado da sigla dos atributos da Tabela 7 pode ser encontrado no apêndice E.

ID	ID_BENEF	ID_DATA	AC	AI	AG	A70	C	CV	ID_CID
3	17851	18473	NULL	71741	36069	NULL	NULL	34918	870
5	9686	18520	NULL	71743	36069	NULL	NULL	NULL	870
6	13442	18530	NULL	71741	36061	NULL	NULL	34918	870
7	19274	18541	NULL	71743	36069	NULL	NULL	NULL	874
8	23950	18543	NULL	71743	36061	NULL	NULL	34918	870
10	46163	18578	NULL	71743	36069	NULL	NULL	NULL	874
11	34075	18579	3414	NULL	NULL	NULL	4626	NULL	13027
13	2957	18584	NULL	NULL	36062	NULL	NULL	34919	2108
14	10534	18584	3414	NULL	NULL	NULL	4626	NULL	13027
15	15012	18584	3414	NULL	NULL	NULL	4626	NULL	13027
16	27399	18584	NULL	NULL	36059	NULL	4626	NULL	10251
17	31554	18584	3414	NULL	NULL	NULL	4626	NULL	13027
18	41818	18584	NULL	71743	36069	NULL	NULL	NULL	874
19	12853	18585	NULL	NULL	36069	NULL	NULL	34919	1859
20	41818	18585	NULL	71743	NULL	NULL	NULL	34919	874
21	47196	18585	3414	NULL	NULL	NULL	4626	NULL	13027
22	2141	18586	NULL	NULL	NULL	49414	NULL	NULL	7152
23	41818	18586	NULL	71743	36069	NULL	NULL	NULL	874
24	2590	18587	3414	NULL	36062	NULL	NULL	NULL	7218
25	6617	18587	NULL	71743	36069	NULL	NULL	NULL	870
26	38541	18587	NULL	NULL	NULL	NULL	NULL	NULL	7142
27	21282	18588	NULL	NULL	36059	49414	NULL	34919	1414
28	11680	18589	NULL	NULL	36069	49414	NULL	34919	1414
29	26411	18589	3414	NULL	NULL	NULL	4626	NULL	13027
30	33929	18589	NULL	NULL	36062	NULL	NULL	34919	13071
31	41154	18589	NULL	NULL	36069	49414	NULL	34919	1414
32	2251	18590	NULL	NULL	36069	49414	NULL	34919	1414
33	2662	18590	NULL	71743	36069	NULL	NULL	NULL	2682
34	6325	18590	NULL	NULL	36069	49414	NULL	NULL	1414
35	23937	18590	3414	NULL	NULL	NULL	4626	NULL	13027
36	27124	18590	3414	NULL	NULL	NULL	4627	NULL	13027
38	13873	18591	3414	NULL	NULL	NULL	4626	NULL	13027
39	41154	18591	NULL	NULL	36061	49414	NULL	NULL	9804
40	21163	18592	NULL	NULL	36069	49414	NULL	NULL	1414
41	25295	18592	3414	NULL	NULL	NULL	4626	NULL	13027
42	6253	18593	NULL	NULL	NULL	NULL	NULL	34917	4358
43	18666	18593	3414	NULL	NULL	NULL	4627	NULL	13027
44	29702	18593	NULL	NULL	36069	49414	NULL	NULL	1414
45	864	18594	NULL	NULL	36061	NULL	NULL	34917	4772
46	21133	18594	3414	NULL	NULL	NULL	4626	NULL	13027
47	6772	18595	NULL	NULL	36069	49414	NULL	NULL	11402
48	35857	18595	NULL	NULL	36061	49414	NULL	NULL	1414
49	38004	18595	NULL	29625	36069	NULL	NULL	NULL	874
50	14847	18596	NULL	NULL	36062	NULL	NULL	34919	7694
52	26023	18596	NULL	71741	NULL	80569	NULL	NULL	2986
56	17202	18599	NULL	NULL	36069	NULL	NULL	34919	2121
58	27064	18599	NULL	NULL	36062	NULL	NULL	34919	13027

Tabela 7: Tabela `FATO_EVENTO_GRUPO`

Como a Tabela 7 é resultante do processo de preparação de dados está exibindo somente os identificadores dos eventos e das CID's, apresenta-se então a Tabela 8 que representa a descrição de cada um dos indicadores dos eventos da Tabela 7, permitindo assim uma visão mais clara da tabela resultante. O significado da sigla dos registros dessa tabela está contido no apêndice F.

ID	ID_BENEF	ID_DATA	AC	AI	AG	A70	C	CV	ID_CID
3	17851	24/04/2011	NULL	100x10VDHY	AG40X12	NULL	NULL	CV20	ABAG
5	9686	14/09/2010	NULL	100x10PLHY	AG40X12	NULL	NULL	NULL	ABAG
6	13442	24/09/2010	NULL	100x10VDHY	AG25X07	NULL	NULL	CV20	ABAG
7	19274	05/10/2010	NULL	100x10PLHY	AG40X12	NULL	NULL	NULL	OTDAB
8	23950	07/10/2010	NULL	100x10PLHY	AG25X07	NULL	NULL	CV20	ABAG
10	46163	11/11/2010	NULL	100x10PLHY	AG40X12	NULL	NULL	NULL	OTDAB
11	34075	12/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
13	2957	17/11/2010	NULL	NULL	AG25X08	NULL	NULL	CV20	EMG
14	10534	17/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
15	15012	17/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
16	27399	17/11/2010	NULL	NULL	AG13X45	NULL	C	NULL	HTABTA
17	31554	17/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
18	41818	17/11/2010	NULL	100x10PLHY	AG40X12	NULL	NULL	NULL	OTDAB
19	12853	18/11/2010	NULL	NULL	AG40X12	NULL	NULL	CV22	NEOPL
20	41818	18/11/2010	NULL	100x10PLHY	NULL	NULL	NULL	CV22	OTDAB
21	47196	18/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
22	2141	19/11/2010	NULL	NULL	NULL	50MLRIO	NULL	NULL	FML
23	41818	19/11/2010	NULL	100x10PLHY	AG40X12	NULL	NULL	NULL	OTDAB
24	2590	20/11/2010	ACUO	NULL	AG25X08	NULL	NULL	NULL	EDT
25	6617	20/11/2010	NULL	100x10PLHY	AG40X12	NULL	NULL	NULL	ABAG
26	38541	20/11/2010	NULL	NULL	NULL	NULL	NULL	NULL	FEJ
27	21282	21/11/2010	NULL	NULL	AG13X45	50MLRIO	NULL	CV22	PRN
28	11680	22/11/2010	NULL	NULL	AG40X12	50MLRIO	NULL	CV22	PRN
29	26411	22/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
30	33929	22/11/2010	NULL	NULL	AG25X08	NULL	NULL	CV22	EOOE
31	41154	22/11/2010	NULL	NULL	AG40X12	50MLRIO	NULL	CV22	PRN
32	2251	23/11/2010	NULL	NULL	AG40X12	50MLRIO	NULL	CV22	PRN
33	2662	23/11/2010	NULL	100x10PLHY	AG40X12	NULL	NULL	NULL	LIPNC
34	6325	23/11/2010	NULL	NULL	AG40X12	50MLRIO	NULL	NULL	PRN
35	23937	23/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
36	27124	23/11/2010	ACUO	NULL	NULL	NULL	CI	NULL	EMG
38	13873	24/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
39	41154	24/11/2010	NULL	NULL	AG25X07	50MLRIO	NULL	NULL	HEMIG
40	21163	25/11/2010	NULL	NULL	AG40X12	50MLRIO	NULL	NULL	PRN
41	25295	25/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
42	6253	26/11/2010	NULL	NULL	NULL	NULL	NULL	CV18	EGINV
43	18666	26/11/2010	ACUO	NULL	NULL	NULL	CI	NULL	EMG
44	29702	26/11/2010	NULL	NULL	AG40X12	50MLRIO	NULL	NULL	PRN
45	864	27/11/2010	NULL	NULL	AG25X07	NULL	NULL	CV18	BRPUL
46	21133	27/11/2010	ACUO	NULL	NULL	NULL	C	NULL	EMG
47	6772	28/11/2010	NULL	NULL	AG40X12	50MLRIO	NULL	NULL	COLNEF
48	35857	28/11/2010	NULL	NULL	AG25X07	50MLRIO	NULL	NULL	PRN
49	38004	28/11/2010	NULL	10EQ	AG40X12	NULL	NULL	NULL	OTDAB
50	14847	29/11/2010	NULL	NULL	AG25X08	NULL	NULL	CV22	DABPEL
52	26023	29/11/2010	NULL	100x10VDHY	NULL	1000MLSEG	NULL	NULL	DLOMB
56	17202	02/12/2010	NULL	NULL	AG40X12	NULL	NULL	CV22	MONON
58	27064	02/12/2010	NULL	NULL	AG25X08	NULL	NULL	CV22	EMG

Tabela 8: Tabela `FATO_EVENTO_GRUPO` **- Descrição dos Eventos**

5 CONCLUSÃO

Com base nos resultados alcançados neste trabalho, pode-se concluir que é possível fazer a conversão dos registros dos eventos em atributos de uma tabela, utilizando a técnica de Fayyad et al. (1996), combinada com algoritmo de Levenshein para descobrir a semelhança entre as descrições dos eventos, bem como a criação da ferramenta que de suporte a um especialista da área da saúde, a fazer uma analise mais objetiva podendo assim gerar uma tabela com dados de maior qualidade para utilização de algoritmo de MD para previsão de CID.

Para trabalhos futuros é evidente a necessidade de aperfeiçoar a ferramenta no quesito de usabilidade e performance, e ou utilizar outros algoritmos para fazer a comparação de semelhança entre os eventos, foi detectado também ao desenvolver o algoritmo que cria o *script* SQL para fazer a inserção dos dados na tabela `FATO_EVENTO_GRUPO` que quando ocorre a situação de um evento do mesmo grupo em um mesmo atendimento do beneficiário o algoritmo não sabe qual evento será escolhido, uma possível solução seria a criação de uma hierarquia entre os eventos, escolhendo então nesses casos o evento de maior prioridade.

Também fica para uma próxima etapa a implementação da aplicação para trabalhar apenas com banco de dados, eliminando assim mais uma dificuldade na utilização da ferramenta que é a leitura e gravação de informação em arquivos do tipo CSV.

A tela de eliminação de redundância ser implementado opção que não permite que o usuário consiga colocar mais de um grupo para o evento selecionado na tela, os outros registros listados devem servir apenas para visualização do usuário, pois da forma como está escrito o algoritmo que gera o *script* SQL de inserção dos dados no `DWSAUDE`, não vai conseguir selecionar o grupo correto.

REFERÊNCIAS

ALVES, A.; GUEDES, A.; REVOREDO, K.; BAIAO, F. Uma metodologia para o aprendizado de um modelo classificador para o alinhamento de ontologias. **Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 doi: 10.5329/RESI**, v. 12, n. 3, 2013.

BARROS, E. F.; ROMÃO, W.; CONSTANTINO, A. A.; SOUZA, C. L. de. Pré-processamento para mineração de dados sobre beneficiários de planos de saúde suplementar. **Journal of Health Informatics**, v. 3, n. 1, 2011.

BRASIL, L. 9961, de 28 de janeiro de 2000. **Cria a Agência Nacional de Saúde Suplementar – ANS e dá outras providências. Disponível na Internet via http://www.planalto.gov.br/ccivil_03/leis/L9961.htm Acesso em 10/06/2015**, 2000.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. knowledge discovery and data mining towards a unifying framework. 1996.

GONDIM F, M. **Algoritmo de Comparação de Strings para Integração de Esquemas de Dados**. 2006. 14 p.

HAN, J.; KAMBER, M.; PEI, J. Data mining: Concepts and techniques. Morgan Kaufmann, 2006.

KOUYLEKOV, M.; MAGNINI, B. Recognizing textual entailment with tree edit distance algorithms. **Recognizing Textual Entailment**, p. 17–20, 2005.

MÉDICA, C. de É. Resolução cfm nº 1931/2009. **Capítulo IX, Art.73. Disponível em: http://www.portalmedico.org.br/resolucoes/cfm/2009/1931_2009.htm. Acesso em 10/06/2015**, 2009.

REZENDE, S. **Sistemas Inteligentes: fundamentos e aplicações. Editora Manole Ltda**. 2003.

RIAÑO, D. Automatic construction of descriptive rules. **AI Communications**, IOS Press, v. 11, n. 1, p. 75–76, 1997.

SILVA, B. J. V. da; MOTTA, R.; LOPES, A. de A. Detecção de autores duplicados utilizando estrutura de comunidades em redes de cooperação científica. 2011.

APÊNDICE A – DESCRIÇÃO DOS ATRIBUTOS DA TABELA 2 - FATO_EVENTO_GRUPO

- ID_BENEF - Identificador do beneficiário
- ID_DATA - Identificador de data
- GE1 - Grupo Evento 1
- GE2 - Grupo Evento 2
- GE3 - Grupo Evento 3
- GEN - Grupo Evento N
- ID_CID - Identificador de CID

APÊNDICE B – DESCRIÇÃO DOS REGISTROS DA TABELA 4 - ARQUIVO COMPARACAOTRATADAID.CSV

- LEPT - LEPTOSPIROSE,AGLUTINACAO(MACROEMICROSCOPIA)
- LINFH - LINFOCITOSTHELPER,CONTAGEMDE(IFICOMOKT-4)(CD4)CITOMETRIADEFLUXO
- LINFS - LINFOCITOSTSUPRESSORES,CONTAGEM(IFCOMOKT-8)(CD8)CITOMETRIADEFLUXO
- LIST - LISTERIOSE-AGLUTINACAO,PORANTIGENO
- MONON - MONONUCLEOSEMONOTESTE
- PROT - PROTEINACREATIVADETERMINACAOQUANTITA-TIVA(TURBIDIMETRIAOUNEFELOMETRIA)
- PROTP - PROTEINACREATIVA,PESQUISA

APÊNDICE C – DESCRIÇÕES DOS REGISTROS DA TABELA 5 - ARQUIVO MENORSTRDIVIDIDAID2.CSV

- LINFH - LINFOCITOSTHELPER,CONTAGEMDE(IFICOMOKT-4)(CD4)
- LINFS - LINFOCITOSTSUPRESSORES,CONTAGEM(IFCOMOKT-8)
- PROT - PROTEINACREATIVA,PESQUISA
- PROTJONES - PROTEINASDEBENCEJONES,PESQUISA
- RUBESIGM - RUBEOLAANTICORPOSIGM,ELISA
- RUBESIGG - RUBEOLAANTICORPOSIGG,ELISA
- RUBEASIGG - RUBEOLA-ANTICORPOSIGG
- RUBEASIGM - RUBEOLA-ANTICORPOSIGM
- RUBESIGG - RUBEOLAANTICORPOSIGG,ELISA
- SIFIGG - SIFILIS(FTA-ABS)IGG
- SIFABS - SIFILISFTAABS
- SIFIGG - SIFILIS-FTA-ABS-IGG
- SIFIGM - SIFILIS-FTA-ABS-IGM

APÊNDICE D – DESCRIÇÕES DO MENU FERRAMENTAS DA FIGURA 9

•Criar Banco de Dados: Apenas cria o banco de dados BDSAUDE vazio.

•Gerar SQL *Update* FATO_EVENTO: Gera um *script* que permite unificar todas os eventos no DWSAUDE que foram marcados para juntar, na seção 3.3.

•Gerar SQL Criação Tabela TGE_GRUPOS – DWSAUDE: O DW DWSAUDE irá ganhar uma nova tabela contendo o domínio dos eventos, a ferramenta já disponibilizará o *script* SQL para aplicar a inclusão da tabela TGE_GRUPOS.

•Gerar Inserção Dados Tabela TGE_GRUPOS – DWSAUDE: Gera o *script* contendo os registros a serem incluídos na tabela TGE_GRUPOS.

•Gerar SQL Criação Tabela FATO_EVENTO_GRUPO – DWSAUDE : *Script* de criação da tabela FATO_EVENTO_GRUPO.

•Gerar SQL Inserção Dados Tabela FATO_EVENTO_GRUPO - DWSAUDE: Gera um *script* contendo os registros a serem incluídos na tabela FATO_EVENTO_GRUPO.

APÊNDICE E – DESCRIÇÕES DOS ATRIBUTOS DA TABELA 7 - FATO_EVENTO_GRUPO

- ID_BENF - Identificador do beneficiário
- ID_DATA - Identificador de data
- AC - Ácidos
- AI - Aguá para Injeção
- AG - Agulhas
- A70 - Álcool 70%
- C - Cálcio
- CV - Cateter Venoso
- ID_CID - Identificado Classificação Internacional de Doença

APÊNDICE F – DESCRIÇÃO DOS REGISTROS DA TABELA 8 - FATO_EVENTO_GRUPO - DESCRIÇÃO DOS EVENTOS

- ACUO - ACIDO URICO
- 100x10VDHY - AGUA P/ INJECAO - Cx. 100 x 10 ml vidro - HYPOFARMA
- 100x10PLHY - AGUA P/ INJECAO - Cx. 100 x 10 ml plast. - HYPOFARMA
- 10EQ - AGUA PARA INJECAO - 10 ml - EQUIPLEX
- AG40X12 - AGULHA DESCARTAVEL 40 X 12
- AG25X07 - AGULHA DESCARTAVEL 25 X 07
- AG25X08 - AGULHA DESCARTAVEL 25 X 08
- AG13X45 - AGULHA DESCARTAVEL 13 X 4 5
- 50MLRIO - ALCOOL 70
- 1000MLSEG - ALCOOLABOR 70 (Restrito Hosp.) - 1000 ml - SEGMENTA
- C - CALCIO
- CI - CALCIO IONIZAVEL
- CV20 - CATETER VENOSO PERIFERICO NO 20
- CV22 - CATETER VENOSO PERIFERICO NO 22
- CV18 - CATETER VENOSO PERIFERICO NO 18
- ABAG - ABDOME AGUDO
- OTDAB - OUTRAS DORES ABDOMINAIS E AS NÃO ESPECIFICADAS
- EMG - EXAME MÉDICO GERAL

- HTABTA - HIPOTERMIA NÃO ASSOCIADA À BAIXA TEMPERATURA AMBIENTAL
- NEOPL - NEOPL BENIG DO COLON RETO CANAL ANAL E ANUS
- FML - FRATURA DO MALÉOLO LATERAL
- EDT - ENTORSE E DISTENSÃO DO TORNOZELO
- FE - FERIMENTO DO JOELHO
- PRN - PROBLEMA NÃO ESPECIFICADO RELACIONADO COM FACILIDADES MÉDICAS E COM OUTROS CUIDADOS DE SAÚDE
- EOOE - EXAME E OBSERVAÇÃO POR OUTRAS RAZÕES ESPECIFICADAS
- LIPNC - LIPODISTROFIA NÃO CLASSIFICADA EM OUTRA PARTE
- HEMIG - HEMORRAGIA DO INÍCIO DA GRAVIDEZ, NÃO ESPECIFICADA
- EGINV - EXAME GERAL INVEST PESS S/QUEIX DIAGN RELAT
- BRPUL - BRÔNQUIOS OU PULMÕES, NÃO ESPECIFICADO
- COLNEF - CÓLICA NEFRÉTICA NÃO ESPECIFICADA
- DABPEL - DOR ABDOMINAL E PELVICA
- DLOMB - DOR LOMBAR BAIXA
- MONON - MONONEUROPATIAS DOS MEMBROS SUPER
- BEN - BENZENO
- IVNE - INFECÇÃO VIRAL NÃO ESPECIFICADA