RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR

ABORDAGEM DE PREPARAÇÃO DE DADOS EM ESCOPO DE REVISTA UTILIZANDO CONJUNTOS APROXIMATIVOS

DATA PREPARATION APPROACH IN MAGAZINE SCOPE USING APPROXIMATIVE SETS

Gustavo Damião Magina de Oliveira[1], Marcos Vinícius Sadala Barreto[2]

Submetido em: 04/04/2021
Aprovado em: 24/04/2021

RESUMO
O projeto MBYP (Magazine Better to Your Paper), que tem por objetivo desenvolver uma plataforma para pesquisadores com o intuito de reduzir o tempo de espera na escolha da revista por meio de métodos de aprendizagem de máquina, surgiu por meio da dificuldade de pesquisadores de compreender se o seu trabalho é adequado aos critérios indexados ao periódico pretendido. Assim, caso não seja, os pesquisadores acabam tendo boa parte do seu tempo perdido e, consequentemente, isso acaba levando a uma perda de confiança quanto ao trabalho desenvolvido. Por isso foi usada a plataforma Sucupira, a fim de se obter dados relacionados aos periódicos nela mantidos para que fossem estudados. Entretanto, inconsistências de dados e elementos duplicados foram encontrados durante a análise das informações obtidas. Devido a isso, foi utilizada a Teoria de Conjuntos Aproximativos (TCA) para eliminar as entidades duvidosas da base de dados principal por meio de procedimento durante o processo de inserção das tuplas no banco de dados construído no MySQL Workbench, além disso, foram criadas funções que aplicam as métricas pertencentes ao TCA para medir o grau de imprecisão dos elementos adquiridos.

PALAVRAS-CHAVE: Teoria dos Conjuntos Aproximativos (TCA). Plataforma Sucupira. Reengenharia. Banco de Dados.

ABSTRACT
The MBYP project (Magazine Better to Your Paper), which purposes to develop a platform for researchers in order to reduce the waiting time to make the choice of the magazine through machine learning methods, has arisen because the difficulty of researchers to understand if their work is adequate to the criteria indexed to the intended journal. Thus, in case it is not, they end up having a good part of their dedicated time lost and consequently ends up leading to a loss of confidence in the work developed. For this, the Sucupira platform was used, in order to obtain data related to the journals kept there so that they could be studied. However, data inconsistencies and duplicate elements were found during the analysis of the information obtained. Because of this, it was used the Rough Set Theory (RST) to eliminate doubtful entities from the main database through a procedure during the tuple insertion process in the database built in the MySQL Workbench. After that, functions were created that apply the metrics belonging to the RST to measure the degree of inaccuracy of the elements acquired.

KEYWORDS: *Theory of Rough Sets (TRS). Sucupira Platform. Reengineering. Database.*

[1] Tecnólogo em Analise e Desenvolvimento de Sistema, IFPA
[2] Doutorado em Engenharia Elétrica pela Universidade Federal do Pará (2019) na área de computação aplicada. Atualmente é professor Educação Básica, Técnico e Tecnológico do Instituto Federal de Educação, Ciência e Tecnologia do Pará atuante nos cursos de Tecnólogo em Processamento de Dados e Técnico em Informática. Tem experiência na área análise de sistemas, persistência poliglota, modelagem matemática, controle retroalimentado, ciência de dados e auditoria em sistemas.

INTRODUÇÃO

Ao longo de toda a história da humanidade, os seres humanos passaram por muitas evoluções até chegarem ao momento atual. Nos primórdios, os hominídeos usavam-se de gritos, gestos, símbolos (pinturas rupestres) e sinais para manter a comunicação com os grupos em que viviam, mas, devido à dificuldade em memorizar gestos, sons e outros sinais, não era possível a criação de uma cultura mais complexa como a fala (11). Milênios mais tarde e após inúmeros processos da evolução, os sumérios desenvolveram um sistema de palavras sistematizadas, vindo a ser chamado posteriormente de escrita cuneiforme, o que permitiu uma comunicação mais avançada entre os povos (12).

Segundo (6), a invenção da escrita contribuiu muito para o desenvolvimento da civilização, pois permitiu atender a um anseio importante: o de registrar o conhecimento. Sendo assim, o armazenamento de dados era de suma importância para a evolução humana.

Assim, após muitos anos de mudanças no mundo, surgiu a computação, e com ela muitos anos depois, entre os anos de 1950 e início de 1960, nasceu a linguagem de programação, nomeada COBOL (Common Business Oriented Language) que tinha por objetivo servir como uma linguagem de programação para negócios, capaz de fazer C.R.U.D (Create, Remove, Update, Delete) em bases de dados e operações aritméticas (14). E com ela surgiu o SGBD (Sistema Gerenciador de Banco de Dados), cujo propósito era retirar do cliente a responsabilidade de gerenciar o acesso, a persistência, a manipulação e a organização dos dados (13). Contudo, após esta nova descoberta, surgiram vários outros problemas, e dentre eles um dos maiores seriam os dados inconsistentes.

A inconsistência de dados é um dos problemas que torna uma base de dados inutilizável. De acordo com (17): “Um banco de dados que se encontre em um estado inconsistente tem a possibilidade de fornecer informações incorretas ou contraditórias a seus usuários”. Para resolver tais problemas, existem várias abordagens que podem afetar uma única tupla, ou o conjunto inteiro de uma tupla, ou até o esquema que está sendo tratado.

Para este trabalho, será utilizado procedimentos para a busca de inconsistência no esquema extraído da plataforma SUCUPIRA (A Plataforma Sucupira, de acordo com o Manual de preenchimento da Plataforma Sucupira, “É uma nova e importante ferramenta para coletar informações, realizar análises, avaliações e ser a base de referência do Sistema Nacional de Pós-Graduação-SNPG.“)(15) na base de dados 2010-2012 e 2013-2016 (16). Foram utilizados procedimentos de buscas que mostram a efetividade das principais regras de normalização nas entidades e procedimentos por tupla, para verificação da ratificação das informações, isso fazendo por amostragem, inicialmente.

A. Problema

O projeto MBYP (Magazine Better Your Paper) tem por objetivo desenvolver uma ferramenta digital, que proporcione ao pesquisador o indicativo de revistas com uma maior probabilidade de aceitação do seu trabalho quanto ao tema desenvolvido, minimizando o tempo de latência para a publicação.

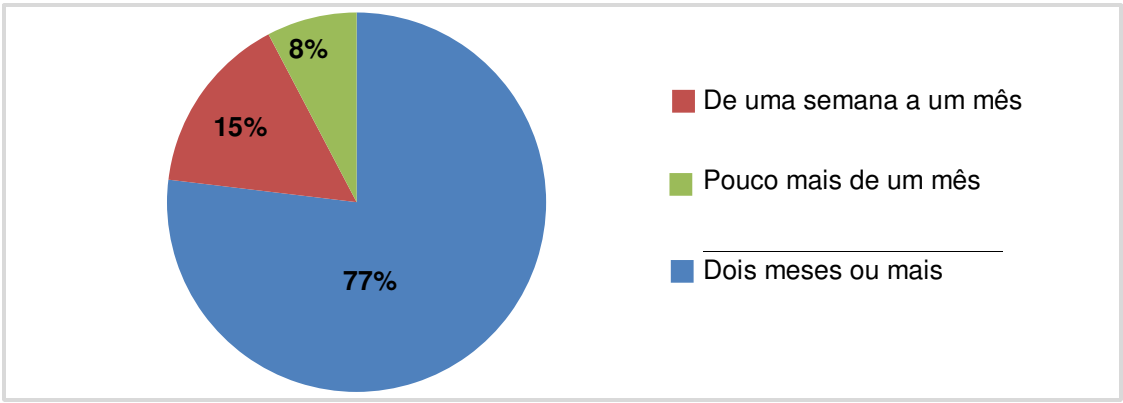
Para que um artigo seja aceito em uma revista há uma série de etapas. O foco deste trabalho está no atendimento dos critérios do escopo da revista, no qual a tal informa ao pesquisador se o tema específico do seu trabalho é adequado ao periódico indexado pretendido. Nesta etapa, não há retorno ao pesquisador informando o porquê de não ser adequado, o que possibilitaria ajustes e correções no trabalho por parte do pesquisador.

Assim, o tempo de preparar o trabalho ao formato do periódico somado ao tempo para o envio mais o tempo de espera da resposta do periódico são desperdiçados, trazendo prejuízos materiais, financeiros e, o mais importante, diminuindo a confiança da equipe, pesquisadores, no trabalho desenvolvido.

Desse modo, foi enviado um questionário para pesquisadores CAPES e Professores Pesquisadores, tanto internos do IFPA quanto de outras universidades. O questionário teve o objetivo de corroborar o pensamento de que se tem certa demora na resposta da correção da revista aos pesquisadores e que seria interessante a eles terem uma ferramenta como o MBYP, que os auxiliasse na escolha da revista que melhor atende ao escopo do seu trabalho.

Tal pesquisa levantou que 76,9% participantes afirmaram que o tempo comumente levado pelos corretores para enviar as correções a serem realizadas é de dois meses ou mais; para 7,7% levou pouco mais de um mês e para aproximadamente 15%, de uma semana a um mês, como pode ser visto na figura 1, e com mais atenção sobre essa e outras respostas da pesquisa no (16).

Figura 1. Tempo comumente levado por revisores de revistas para responder pesquisadores



Existem bases de dados que disponibilizam os dados sobre os periódicos indexados, o que auxilia na escolha da revista por diversos outros fatores, p. ex. a plataforma SUCUPIRA, que é mantida pelo Conselho Nacional de Pesquisa – CNPQ – e que tem como o objetivo principal de divulgar o QUALIS dos periódicos. QUALIS é o:

sistema brasileiro de avaliação de periódicos mantido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, que relaciona e classifica os veículos utilizados para a divulgação da produção intelectual dos programas de pós-graduação do tipo stricto sensu, quanto ao âmbito da circulação e à qualidade, por área de avaliação.(1)

Entretanto, essa base de dados em nada auxilia para a resolubilidade da questão principal: qual o escopo da revista? Apesar de ser a principal base de dados utilizada pelos pesquisadores brasileiros vinculados às instituições de ensino, durante os estudos da base de dados, foram encontradas diversas inconsistências, informações duplicados, informações quebradas e elementos divergentes. Uma amostra desses dados pode ser vista no quadro 1.

Quadro 1: Amostra de dados imprecisos obtidos na plataforma Sucupira

ISSN	TITULO	ÁREA	EXTRATO
2257-0543	ARTELOGIE	INTERDISCIPLINAR	B2
2257-0543	Bresil(s)	ANTROPOLOGIA/ARQUEOLOGIA	B3
2238-0167	Revista Jiop	LETRAS / LINGUÍSTICA	B5
2238-0167	Revista da Extensão	EDUCAÇÃO	C
2237-4957	Revista do Núcleo Onetti de Estudos Literários Latino-Americanos	LETRAS / LINGUÍSTICA	C
2237-4957	Guará Linguagem e Literatura	LETRAS / LINGUÍSTICA	B5

Fonte: Base de dados da plataforma Sucupira

No quadro 1, é possível identificar 4 colunas, sendo elas ISSN, TITULO, ÁREA e EXTRATO. O ISSN é classificado como sendo uma chave de identificação única composta por oito caracteres e cada revista deve possuir somente uma chave. Desse modo, a existência de duas revistas com títulos diferentes compartilhando o mesmo ISSN é considerada uma imprecisão nos dados.

Assim, é possível eliminar as inconsistências encontradas nos esquemas extraídos da plataforma SUCUPIRA, nas bases de dados referentes aos anos de 2010 e 2016, de forma automatizada?

B. Justificativa

Durante o dia a dia, somos bombardeados por informações. Desde o momento em que acordamos e pegamos o celular, ligamos a televisão ou computador, recebemos os mais diversos dados: da previsão do clima para o dia a endereços eletrônicos pessoais e de trabalho.

Supondo que uma pessoa comum possua a rotina de acordar cedo e ver a previsão do tempo para tomar a decisão de se vai ou não levar um casaco ao sair, o que poderia acontecer caso a previsão do clima estivesse errada? A pessoa hipotética iria se molhar na chuva ou iria carregar o casaco sem necessidade. A mesma situação pode ser espelhada para inúmeras outras situações: no lugar do casaco poderia ser uma pasta protetora para documentos jurídicos, eletrônicos ou afins.

Saindo de uma situação hipotética, pode-se citar relatos que caracterizam a hipótese em realidade, a citar:

> Colaboradores com muito tempo de casa ou a falta de busca constante da atualização dos dados dos colaboradores podem gerar informações divergentes que, em um processo de saque de FGTS, pode trazer muita dor de cabeça – evitada somente com um bom saneamento da base de dados(29).

Relatos como esse são comuns de acontecer quando se usa informações incertas. Outro exemplo vem do meio comercial: a informação é o ponto-chave do negócio, saber onde os clientes estão, quais os produtos são mais vendidos e em qual período do ano ocorrem mais compras é imprescindível para qualquer empresa do varejo ou do atacado se destacar atualmente no mercado.

Sendo assim, as informações imprecisas são capazes de inutilizar qualquer base de dados que essas empresas possam usar, ocasionando perdas no valor investido nessas pesquisas de dados, pois, segundo (29) “A baixa qualidade dos dados levará a uma baixa qualidade no resultado”.

Mas, o que viriam ser as inconsistências ou imprecisões em dados? De acordo com os dicionários da Língua Portuguesa, “inconsistência” define-se como “Falta de firmeza ou de solidez; inconstância; incerteza” (30); já “dados” podem ser definidos como sinônimo de: conhecimentos, saberes, informações (31). Sendo assim, a “inconsistência de dados” pode ser definida um conjunto de informações com falta de solidez ou falta de certeza. Dessa forma, é interessante compreender que, quando falamos desses dados, estamos falando de elementos com erros, o que tira boa parte de sua credibilidade profissional e científica.

Tais erros existem no mundo da informática desde o surgimento do ENIAC, primeiro computador já criado, e persistem até a atualidade. Dependendo do conteúdo que está sendo estudado em uma pesquisa acadêmica ou profissional, é possível encontrar tal irregularidade quase que constantemente. Na área de *data mining* (processo de explorar abundantes quantidades de dados), é quase rotina a tarefa de livrar bases de conhecimento de tais problemas, como informações quebradas, duplicadas, diferentes (nos casos que deveriam ser idênticas), ruídos, etc.

Sabendo disso, é importante compreender o que pode gerar tais complexidades. Na área de *data mining* (Mineração de dados, em português), geralmente, o trabalho é colocado em algumas etapas, sendo a primeira a coleta de dados. Essa coleta pode ocorrer em bancos de dados públicos, em documentos privados, redes sociais, etc. No caso das redes sociais, naturalmente, algumas informações podem vir com essas falhas, não só pela diversidade sociocultural do país, mas também pela informalidade das publicações. Em outros casos, erros durante a compilação das informações, durante o “download” e/ou o “upload” dos dados e, mais comumente, erros durante a leitura do

arquivo (devido à incompatibilidade da versão do arquivo com a versão do software usado para a leitura).

De acordo com (28),

> Frequentemente, os dados são encontrados com diversas inconsistências: registros incompletos, valores errados e dados inconsistentes. A etapa de limpeza dos dados visa eliminar estes problemas de modo que eles não influam no resultado dos algoritmos usados.

Desse modo, uma limpeza de dados é de suma importância para aprimorar a qualidade da base de conhecimentos brutos que está sendo usada. Existem diversas técnicas que podem ser usadas na limpeza de dados.

O projeto MBYP investigou a base de dados da plataforma Sucupira e encontrou inconsistências em sua base de dados, sendo necessário proceder a uma etapa de limpeza dos dados a fim de se exaurir as incertezas existentes. Assim, para efetuar a retirada das inconsistências encontradas nas bases de dados extraídas da plataforma SUCUPIRA, será utilizada a teoria dos conjuntos aproximativos, um modelo matemático baseado nas regras matemáticas, que faz o uso de conjuntos comumente denominados superior e inferior.

A teoria de conjuntos aproximativos foi criada em 1982 por Pawlak e tem por objetivo o tratamento de imprecisão/inconsistências de dados e redução de tabela, obtendo um resultado muito superior ao obtido nos dados puros, além de levar a um ganho de processamento computacional devido à redução das informações. Nas referências bibliográficas investigadas (2)(7)(8)(9), notou-se a ausência da utilização da teoria de conjuntos aproximativos em modelos de persistência de dados relacionais, mostrando a importância da realização de pesquisas sobre esta ótica.

C. Objetivo Geral

O objetivo do projeto consiste em implementar o algoritmo da Teoria dos Conjuntos Aproximativos e aplicá-lo em um banco de dados de modelo relacional, para resolver imprecisões de revistas.

D. Objetivos Específicos

- Adotar procedimento para exclusão dos atributos de inconsistências na base de dados da plataforma Sucupira;
- Adotar funções para utilizar os coeficientes de aproximações;
- Medir o grau de precisão dos termos imprecisos;

E. Organização do Trabalho

Inicialmente será investigada as fundamentações teóricas que são pertinentes ao desenvolvimento do procedimento autônomo de minimização das incertezas das revistas durante a inserção na base principal do projeto MBYP.

Na fundamentação teórica, é abordada a Teoria dos Conjuntos Aproximativos, tema principal do desenvolvimento deste trabalho, a partir de seus conceitos e características principais: Relações Indiscerníveis, Aproximação dos Conjuntos, Qualidade das Aproximações e Reduções, que serão usados como lógica para destacar elementos inconclusivos e duplicados, além de lhes aplicar métricas para se ter uma melhor compreensão do grau de inconsistência.

Em ambiente de coleta de dados é abordado o ambiente da plataforma Sucupira e os passos realizados para a obtenção das informações, uma breve explanação a respeito dos dados contidos no documento obtido.

Em prototipação de *software*, é abordado como será realizada a análise das informações e o que será considerado na hora de construir a base de dados, além de mostrar sua construção, sua implantação do algoritmo de redução e sua aplicação das métricas.

Em resultados, mostra-se a aplicação da teoria estudada e uma análise dos efeitos da teoria de conjuntos aproximados na persistência de dados do ecossistema estudado.

Em conclusão avaliação, trabalhos futuros e dificuldades encontradas durante o desenvolvimento.

1 FUNDAMENTAÇÃO TEÓRICA

Neste tópico será apresentado a Teoria de Conjuntos Aproximativos utilizada no desenvolvimento deste trabalho.

A. Teoria dos Conjuntos Aproximativos

A Teoria dos Conjuntos Aproximados (TCA) é um modelo matemático proposto pelo polonês Zdzislaw Pawlak para a resolução de alguns problemas, como o tratamento de incerteza e a classificação aproximada (2). De uma maneira simplificada, a TCA pode ser considerada como destinada ao tratamento de inconsistências e imprecisão de dados, através de conjuntos denominados superior e inferior, assim se pode obter um resultado aproximativo e, consequentemente, diminuir a quantidade de atributos, levando a um ganho computacional.

Esta teoria pretende obter relações indiscerníveis, “que diz que dois elementos são ditos indiscerníveis, se possuem as mesmas propriedades, segundo Leibniz” (7). Ou seja, conjuntos que tenham objetos com as mesmas propriedades ou valores que sejam idênticos ou similares.

Os conjuntos da TCA não podem ser caracterizados/definidos exatamente como função do conjunto de atributos disponíveis. Além disso, ele não necessita de nenhuma informação adicional a respeito de dados, como: distribuição de probabilidade, atribuição de crenças, grau de pertinência ou probabilidade.

Dentro da Teoria dos Conjuntos Aproximativos é possível encontrar assuntos relacionados a aproximação de conjuntos e a espaços aproximados, sendo este último nada mais que um par ordenado A = (U,R), em que: U é um conjunto não vazio, denominado conjunto universo, e R é uma relação de equivalência sobre U, denominada Relação de Indiscernibilidade (38).

De acordo com (38), baseando-se nas propriedades de relações binárias, a relação $R \subseteq XxX$ pode ser definida como uma relação de equivalência devido se ligar as propriedades:

- Reflexiva, que nada mais é do que um elemento que está relacionado com ele próprio xRx;
- Simétrica, onde se xRx' , então $x'Rx$; e
- Transitiva, que diz, se xRx' e $x'Rz$, então x*R*z,

Desse modo, os itens $x, x' \in$ U, se xRx' então x e x' são indistinguíveis em S, ou seja, o tipo de correspondência definida por x é a mesma que a definida por x', isto é, $[x]R = [x']R$ (38).

A classe de equivalência de certos elementos $x \in X$ composto por todos os componentes $x' \in X$ para os quais xRx'. Os componentes que são indistinguíveis formam os chamados conjuntos elementares. Desse modo, é possível chegar à conclusão que as classes de equivalência de A são os conjuntos elementares de S (38).

Desse modo, um sistema de informação é um par ordenado S = (U; A), onde U é o universo e A, o atributo.

Cada atributo $a \in A$ é uma função $a: U \rightarrow Va$, em que Va é o conjunto dos valores permitidos para o atributo a (sua faixa de valores) (38).

A representação da abordagem da TCA é mais comumente demonstrada por um sistema de informação que nada mais é do que uma tabela, na qual cada linha representa um caso, evento ou objeto; cada coluna representa um atributo, seja ele uma variável ou propriedade, que pode ser avaliado/medido para cada objeto. E esses atributos são os mesmos para cada um dos objetos; contudo, nominalmente, estes valores podem diferenciar-se (8).

Muitos sistemas de informação utilizam ainda um atributo de decisão. Podendo ser um sistema qualquer de informação na forma de

$$S = (U, A \cup \{d\}) \quad (1)$$

onde *d* é o atributo de decisão.

Os elementos de *A* são chamados atributos condicionais ou simplesmente condições. (7)

O Quadro 2 mostra um exemplo de um pequeno sistema de informação, no qual dez pessoas, representadas por U (Universo), recebem cada uma um conjunto de dados de atributos distintos, e, no final, o atributo de classificação é preenchido em conformidades as regras que representa.

Com base no quadro 2, é possível identificar tais conjuntos dentro do universo exibido (8):

- U = {1,2,3,4,5,6,7,8,9,10} = Conjunto de objetos do universo.
- A = {Tipo, Tamanho, Material, Tema Central, Classificação} = Atributos.
- *d* = {Atitude} = Atributo de Decisão.

Quadro 2: Exemplo de um sistema de informação

U	Atributos					Atributo Decisão
Pessoa	Tipo	Tamanho	Material	Tema Central	Classificação	Atitude
1	RPG	Grande	Plástico	Detetive	Acima de 18 anos	Negativa
2	Labirinto	Médio	Acrílico	Policial	Acima de 10 anos	Neutro
3	Tabuleiro	Pequeno	Papel Laminado	Medieval	Acima de 12 anos	Positiva
4	RPG	Médio	Acrílico	Medieval	Acima de 18 anos	Negativa
5	Tabuleiro	Pequeno	Papel Laminado	Detetive	Acima de 18 anos	Neutro
6	Lego	Grande	Plástico	Policial	Acima de 10 anos	Positiva
7	Tabuleiro	Pequeno	Plástico	Detetive	Acima de 16 anos	Positiva
8	Tabuleiro	Pequeno	Plástico	Detetive	Acima de 18 anos	Positiva
9	Lego	Grande	Plástico	Policial	Acima de 10 anos	Neutro
10	Lego	Médio	Acrílico	Policial	Acima de 18 anos	Neutro

A.1. Relações Indiscerníveis

Quando obtemos uma base de dados, é possível nos deparar com diversas redundâncias, inconsistências, além de muitas informações desnecessárias.

Ponderando que redundâncias podem ser consideradas repetição de informações, os objetos que se enquadram nesse requisito têm as mesmas propriedades e são considerados idênticos ou similares (8).

Levando em consideração os atributos mostrados, se *S =(U,A)*, então para todo subconjunto $B \subseteq A$ existe uma relação de indiscernibilidade $IND_A(B)$, definida como (10):

$$IND_A(B) = \{(x, x') \in U^2\} \forall a \in B, a(x) = a(x')\} \quad (2)$$

Para facilitar o entendimento, o universo do quadro 2 será usado como um exemplo prático. Para isso, será adotado o subconjunto B = {Tipo}, no qual é possível reconhecer objetos indiscerníveis, ou seja, que estão na mesma classe de equivalência, tais como (8): {1,4}, {6,9,10}, {3,5,7,8}.

Cada um dos elementos contidos acima origina uma classe de equivalência correspondente, como é mostrado na Quadro 3.

Quadro 3: EXEMPLO DE UM SISTEMA DE INFORMAÇÃO

Atributos	
Classes	**Tipo**
C11	RPG
C12	Lego
c13	Tabuleiro

A.2. Aproximação dos Conjuntos,

Segundo (7), a relação de igualdade entre as classes pressupõe a participação do universo e essa participação necessitará ser aproveitada para formar subconjuntos do universo. E, nesse momento, será possível observar que alguns elementos não são "definidos corretamente" pelo fato de serem indiscerníveis, o que acabaria provocando decisões incorretas. Dessa forma, estes elementos dividem-se em três classes: os que podem ser classificados pertencentes à classe desejada, os que não pertencem à classe desejada e os que não podem ser classificados. A seguir são expressas formalmente estas noções (9).

Levando em consideração o sistema de informação $S = (U, A)$, e $B \subseteq A$, e $X \subseteq U$, em que é o conjunto de objetos com relação a *B*; desse modo, usando somente as informações dos atributos contidos em *B* é permissível obter *X*. Assim se define Aproximação Inferior de *X* em relação à *B*, indicado por $\underline{B}X$ e Aproximação Superior de *X* em relação a *B*, indicado por $\overline{B}X$ (8)(9)(38), em que:

$$\underline{B}X = \{x \in U | U/INDs(B) \subseteq X\} \quad (3)$$

$$\overline{B}X = \{x \in U | U/INDs(B) \cap X \neq \emptyset\} \quad (4)$$

onde

$\underline{B}X$ = Claramente são membros da classe esperada;

$\overline{B}X$ = Possivelmente são da classe esperada;

RF(x) = $\overline{B}X - \underline{B}X$ = Região de Fronteira = podem não pertencer a tal classe;

$U - \overline{B}X$ = Fora da Região = Não pertencem a classe esperada.

Fazendo novamente uma breve análise no quadro 2, é possível usá-la como exemplo neste caso para uma exemplificação. Para isso, é necessário o questionamento: quais as características dos atributos que definem as atitudes das pessoas com relação aos jogos como sendo Negativa, Neutra ou Positiva? Quando se analisa o quadro 2 novamente sob essa ótica, são identificados dois atributos inconsistentes, 6 e 9, pois eles têm os mesmos atributos, mas possuem o atributo de decisão diferente. Sendo assim, é interessante fazer a aplicação da teoria dos conjuntos aproximativos nestes elementos (8).

- Atitude a ser considerada = Positiva;
- S = {Tipo, Tamanho, Material, Tema Central, Classificação};
- U = {1},{2},{3},{4},{5},{6},{7},{8},{9},{10};

- X = {3,6,7,8} = Elementos Positivos;
- $\overline{B}X$ = {{3},{6,9},{7},{8}}; (Elementos que são positivos com aqueles que talvez sejam positivos).
- $\underline{B}X$ = {{3},{7},{8}}; (Elementos que sem dúvida são positivos).
- RF(x) = $\overline{B}X - \underline{B}X$ = {6,9}; (Elementos duvidosos).
- FR(x) = $U - \overline{B}X$ = {{1},{2},{4},{5},{10}};(Elementos que não são positivos).

Na figura 2 são mostrados os dados do quadro 2 de modo que seja de fácil entendimento os dados sobre aproximação dos conjuntos.

Por meio de uma observação simples na figura 2, é possível identificar que os elementos que fazem parte do grupo selecionado (grupo positivo) são {3}, {7},{8}, definindo assim a aproximação inferior ($\underline{B}X$), como foi destacado na figura 3.

Logo em seguida, são exibidos os elementos da aproximação superior, que pode ser classificada como sendo os elementos da aproximação inferior mais os elementos que podem fazer parte do grupo com o atributo de decisão positiva, nesse caso, {3}, {6,9}, {7}, {8}, como é destacado na figura 4.

Além desses, tem-se os elementos da região de fronteira, que são compostos por entidades classificadas como duvidosas, e no exemplo da figura 5 é possível identificar dois dados, sendo eles 6 e 9. Por fim, é possível destacar os elementos que estão fora da região estudada, como: {1}, {2},{4}, {5}, {10}.

Figura 2: Diagrama de Venn e Teoria de Conjuntos Aproximados

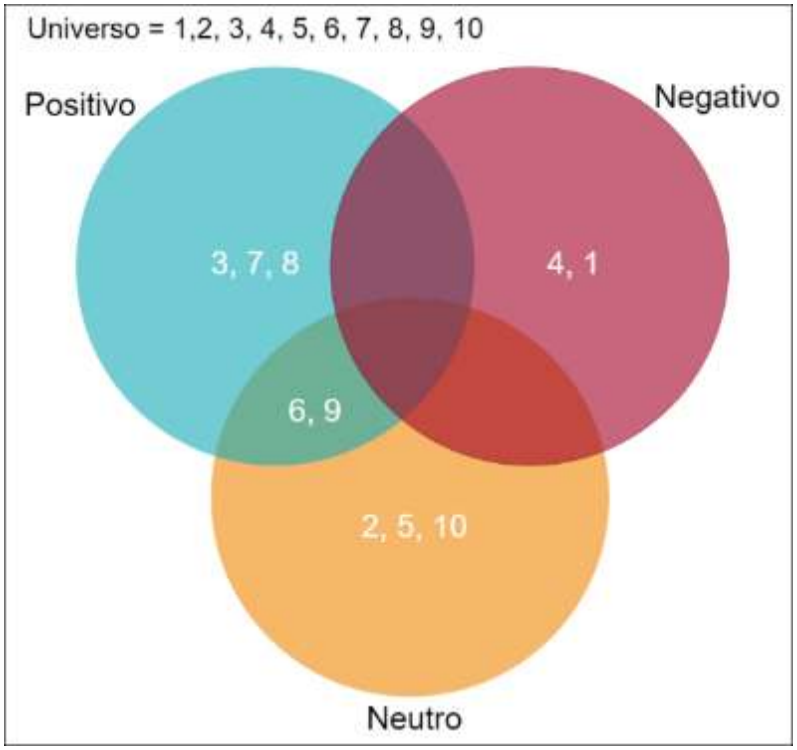


Figura 3: Diagrama de Venn e Aproximação Inferior

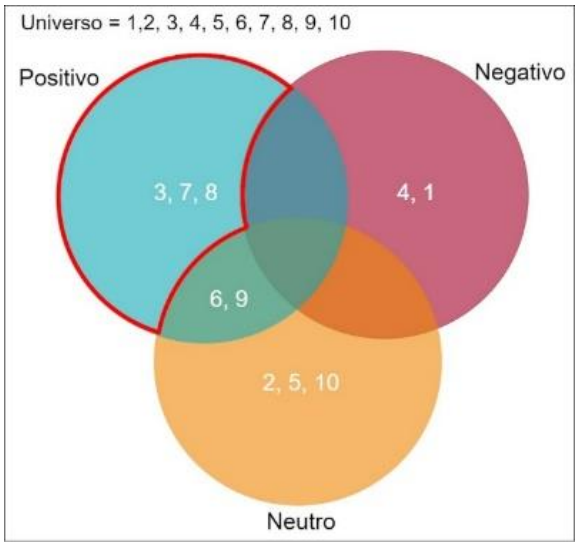


Figura 4. Diagrama de Venn e Aproximação Superior

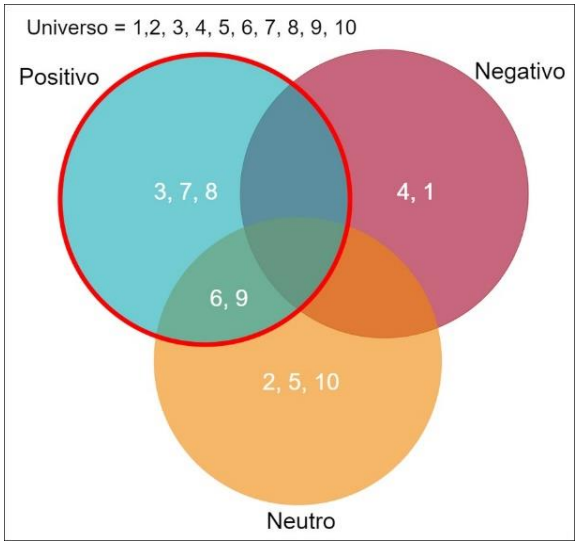
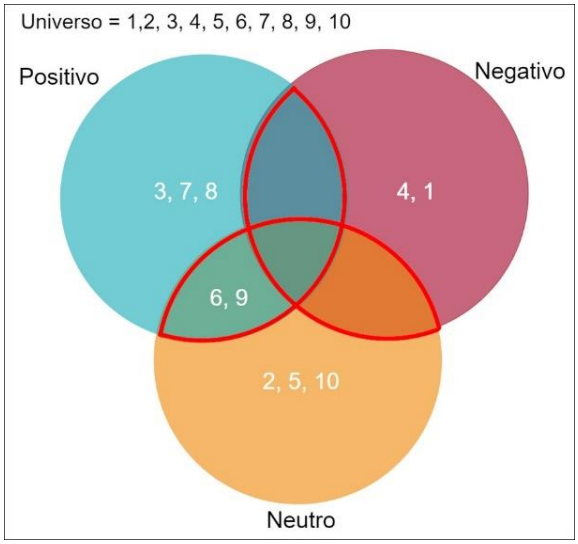


Figura 5. Diagrama de Venn na região de fronteira



A.3. Qualidade das Aproximações

Estes são os coeficientes normalmente utilizados para medir as aproximações e são obtidos por meio do resultado das aproximações superiores e inferiores. Entre os coeficientes é possível destacar o Coeficiente de Imprecisão ($\alpha_B(X)$), Coeficiente da Qualidade da Aproximação Inferior ($\alpha_B(\underline{B}X)$), Coeficiente da Qualidade da Aproximação Superior ($\alpha_B(\overline{B}X)$), segue abaixo as respectivas regras (7):

$$\alpha_b(X) = \frac{|\underline{B}X|}{|\overline{B}X|} \tag{5}$$

$$\alpha_b(\underline{B}X) = \frac{|\underline{B}X|}{|U|} \tag{6}$$

$$\alpha_b(\overline{B}X) = \frac{|\overline{B}X|}{|U|} \tag{7}$$

- Coeficiente de imprecisão: é a qualidade da aproximação de X e segue as seguintes normas:
  - $Se\ 0 \leq \alpha_B \leq 1$, X é dito como impreciso.
  - $Se\ \alpha_B = 1$, X é dito como Preciso.
- Coeficiente da Qualidade da Aproximação Inferior: mede a percentagem de todos os objetos que são relacionados como pertencentes em X.
- Coeficiente da Qualidade da Aproximação Superior: indica o percentual de todos os objetos possivelmente classificados como pertencentes a X.

Baseando-se no quadro 2, podemos lhes aplicar tais métricas da seguinte forma:

Coeficiente de imprecisão = $\frac{[3],[7],[8]}{[3],[6],[9],[7],[8]} = \frac{3}{5} = 0,6$

Coeficiente da Qualidade da Aproximação Inferior =

$$\frac{[3],[7],[8]}{[1],[2],[3],[4],[5],[6],[7],[8],[9],[10]} = \frac{3}{10} = 0,3$$

Coeficiente da Qualidade da Aproximação Superior =

$$\frac{[3],[6],[9],[7],[8]}{[1],[2],[3],[4],[5],[6],[7],[8],[9],[10]} = \frac{5}{10} = 0,5$$

A.4. Reduções

Em (7), simplifica-se as Reduções como sendo nada mais do que a diminuição de redundâncias dentro de um sistema de informação, devido elas provocarem um aumento da dificuldade, ocasionando extrações de informações longas e cansativas.

"Outro artifício para redução é manter somente os atributos que preservam a relação de indiscernibilidade" (7), criando, assim, um conjunto que possui somente as informações mínimas e imprescindíveis para manter a qualidade da classificação se comparado ao conjunto original. Devido a isso os conjuntos mínimos são chamados de reduções.

Ainda de acordo com (9), no sistema de informação $S = (U, A)$, para que se tenha uma redução de *S* em um conjunto mínimo de atributos, através $B \subseteq A$ tal que $INDs(B) = INDs(A)$, ou seja, uma redução (RED(B)) é o conjunto mínimo de atributos de A, que pode ser utilizado preservando o conjunto de atributos do universo completo (7).

Um sistema de informação pode ter mais de uma redução. A intersecção de todas as reduções chamada de núcleo, que é dada por $N(B) = \bigcap RED_i(B), i = 1 \dots n$

2 AMBIENTE DE COLETA DE DADOS

A plataforma Sucupira foi escolhida inicialmente no projeto para a coleta de dados devido à transparência na divulgação das informações, processos e procedimentos realizados pela CAPES para a comunidade acadêmica.

De acordo com (15), a plataforma Sucupira contribuiu para o avanço dos seguintes processos da CAPES:

- Maior transparência dos dados para toda a comunidade acadêmica;
- Redução de tempo, esforços e imprecisões na execução de avaliação do SNPG;
- Maior facilidade no acompanhamento da avaliação;

- Maior confiabilidade, precisão e segurança das informações;
- Controle gerencial mais eficiente.

A seguir, serão abordados aspectos importantes na obtenção, análise e relacionamento entre os dados adquiridos na plataforma em questão.

A. O AMBIENTE

Para conseguir a base de dados, inicialmente, é necessário acessar o link https://sucupira.capes.gov.br/sucupira/, que irá indicar as variáveis ambientais necessárias para o desenvolvimento deste trabalho.

B. AS VARIÁVEIS AMBIENTAIS

Após a obtenção da base de dados, cada coluna obtida foi identificada e analisada para ser realizado o processo de reengenharia. As colunas ISSN, Título, Área de Avaliação e Estrado (16), podem ser traduzidas como sendo:

- ISSN: É uma combinação de oito números, que tem como finalidade reconhecer e especificar o título de uma publicação científica ordenada em âmbito internacional (26).
- TÍTULO: Título do periódico que está sendo pesquisado.
- ÁREA DE AVALIAÇÃO: Trata-se da área na qual um determinado periódico atua. Ex: História, Engenharia, etc.
- ESTRATO: É a classificação do periódico. Ex: C, A1, B1.

Além dessas, para a construção da base de dados, serão utilizadas as variáveis Período, Palavras e Artigos.

- PERÍODO: O período em que o dado foi obtido. Ex: 2010,2016, etc.
- PALAVRAS: Será responsável por armazenar palavras consideradas importantes para determinar o escopo de uma revista. (Seu desenvolvimento será realizado em trabalhos futuros).
- ARTIGOS: Será usada para armazenar os locais onde os artigos estudados estão armazenados.

C. AS REGRAS DE RELACIONAMENTO

De acordo com as variáveis levantadas na página 10, na seção chamada AS VARIÁVEIS AMBIENTAIS, é possível realizar a relação entre os elementos do seguinte modo:

Será criado a tabela REVISTA, que conterá as variáveis Título e ISSN, sendo esta última a chave primária da tabela, devido ao seu conceito de ser um número internacional destinado à identificação de publicações no meio científico, portanto mais apropriada para ser a chave primária da tabela.

A seguir, será criada uma tabela AREA, que conterá os escopos de atuações disponíveis (Ex: Biologia, Geografia, Ciência da Computação, etc.), tal entidade se relacionará com a tabela REVISTA em uma relação de N para N, na qual uma revista pode ter vários escopos e um escopo pode ter várias revistas.

Do mesmo modo, poderá ser previsto para as entidades PALAVRAS e ARTIGOS, que deverão ter um relacionamento entre si, respeitando que cada palavra deverá ter vários artigos e um artigo deverá ter várias palavras, constituindo, assim, uma relação N para N.

Ainda assim, as entidades PALAVRAS e ARTIGOS deverão ser ligadas à tabela REVISTA, e tal ligação respeitará a condição que diz que uma revista terá vários artigos, e um artigo terá uma revista, formando uma relação 1 para N. Enquanto isso, a tabela PALAVRAS possuirá a relação de N para N, devido à relação dizer que cada palavra possui várias revistas, e uma revista possui várias palavras.

Logo adiante, será feita a tabela CLASSIFICAÇÃO, na qual será armazenado o extrato de cada revista. Sua chave primária foi definida como sendo os próprios extratos, já que são únicos (Extratos: A1, A2, B1, B2, B3, B4, B5, C).

Em seguida, será criada a tabela PERIODO, com a finalidade de armazenar o intervalo de tempo a que pertence uma determinada revista na plataforma Sucupira. Sua chave primária serão valores inteiros na ordem de 1 a 2 (até o presente momento), devido terem somente dois intervalos de tempo de escolha na plataforma Sucupira (2010-2012 e 2013-2016).

Tais entidades (CLASSIFICACAO e PERIODO) vão se relacionar na ordem de N para N, uma vez que um período tem várias classificações e uma classificação tem vários períodos.

O resultado das relações entre as tabelas AREA, REVISTA, CLASSIFICAÇÃO e PERIODO será unificado em uma entidade chamada QUALIS, em uma relação de N para N, na qual uma determinada revista com uma determinada área de atuação pode ter várias classificações e períodos, e uma classificação de um determinado período pode ter várias revistas e áreas.

3 PROTOTIPAÇÃO DE *SOFTWARE*

A. REENGENHARIA DA BASE DE DADOS

Após a obtenção da base de dados da plataforma Sucupira, foi discutido como cada um dos elementos contidos na planilha obtida deveriam interagir entre si, dando início a reengenharia da

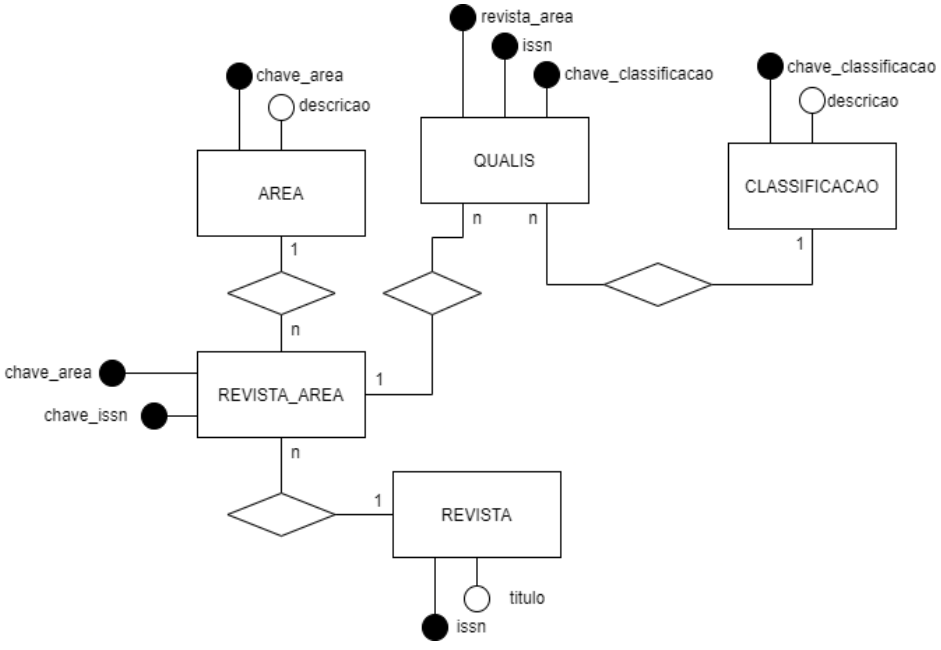
base de dados. Nesse ponto, foi adotado o Modelo Conceitual, de modo a obter um modelo simples e de fácil entendimento da relação entre as entidades, como mostrado no diagrama 13.

Durante o processo de reengenharia da base, cada coluna pertencente a planilha obtida na plataforma Sucupira foi identificada como sendo uma entidade ou um atributo no modelo conceitual. Dentre as três entidades principais, pode-se destacar REVISTA, AREA, CLASSIFICACAO:

- REVISTA: terá dois atributos, ISSN que será a chave primária e TITULO que receberá o respectivo título do ISSN contido na planilha.
- AREA: terá dois Atributos, CHAVE_ÁREA que será a chave primária e DESCRIÇÃO que conterá o nome das áreas em que uma determinada revista realiza publicações.
- CLASSIFICAÇÃO: terá dois atributos, CHAVE_CLASSIFICAÇÃO que será a chave primária e DESCRICÃO que conterá a classificação de uma determinada revista em uma determinada área que poderá ser A1, B1, C, etc. .

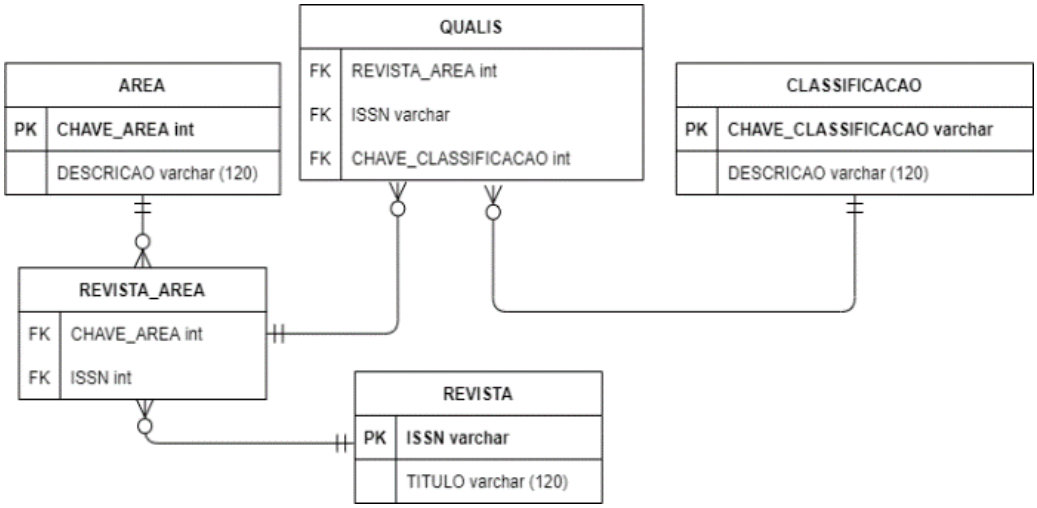
O modelo em questão é observável no diagrama da figura 6, com suas entidades inter-relacionadas.

Figura 6. Reengenharia de *Software* Modelo Conceitual



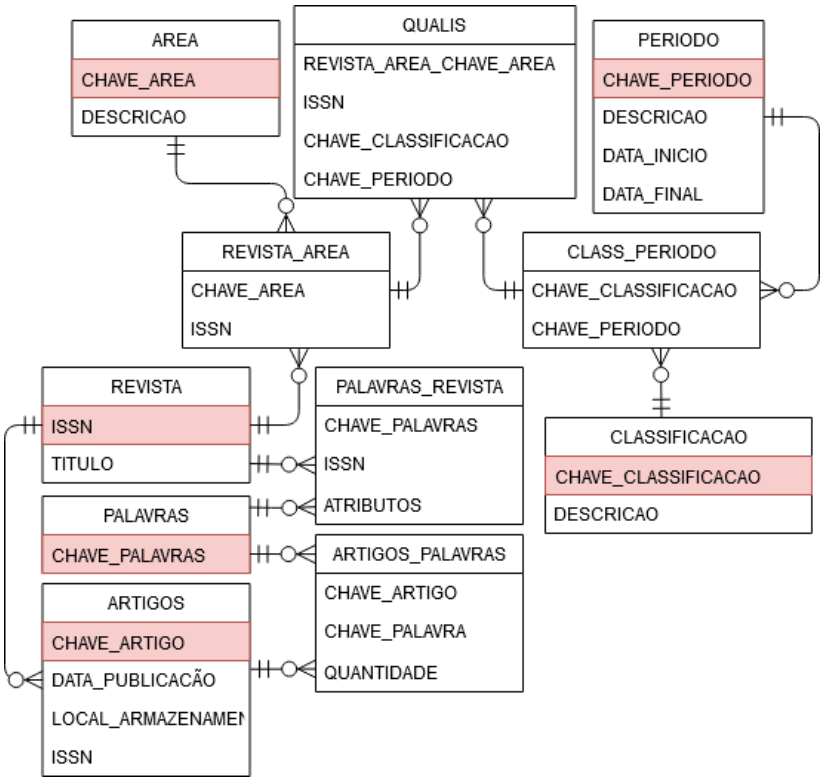
Em seguida, foi feita a modelagem seguindo o Modelo Lógico, acrescentando os tipos que cada campo irá aceitar (Int, Double, Varchar, String, etc.), quantos caracteres cada campo irá permitir e as chaves primárias (PK) e estrangeiras (FK), como mostrado no diagrama da figura 7.

Figura 7. Reengenharia de *Software* Modelo Lógico



Logo após algumas análises mais aprofundadas, foi observado que seria interessante atualizar o modelo relacional normalizado, com o objetivo de atender certas necessidades futuras, como o período no qual as tabelas vinham (2010 ou 2016 até o momento da escrita deste documento), os artigos com suas datas de publicações junto com seu local de armazenamento e outras entidades que serão usadas para trabalhar com mais tecnologias no decorrer do projeto (ARTIGOS_PALAVRAS e PALAVRAS_REVISTAS), como é possível visualizar no diagrama da figura 8.

Figura 8. Modelo Relacional Normalizado



A partir deste ponto foi criado um modelo físico de banco de dados utilizando o MySQL Workbench, como pode ser visto no link do github(16).

E logo em seguida, foi iniciada a preparação dos metadados no SGBD, verificando como cada dado deveria ser posto nas tabelas e qual a prioridade a ser respeitada para o processo de popular as entidades. Assim sendo, foi preferível inserir no banco de dados inicialmente os dados relacionados a tabela “ÁREA”, devido à necessidade de se manter um controle sobre as áreas que serão postas no banco de dados, depois a tabela “PERIODO”, uma vez que seja necessário identificar quando uma determinada revista foi inserida, e, por último, a tabela “CLASSIFICAÇÃO”, já que também é necessário ter um controle sobre os dados que são inseridos, não permitindo, então, que eles sejam colocados na tabela ao mesmo tempo que são colocados outros dados, em que grau de incerteza quanto a sua veracidade é alto.

Logo a seguir, será discutido sobre como será realizado o tratamento de tais imprecisões, quais os procedimentos criados para auxiliar na separação automática de tais tuplas, com uma explicação breve sobre cada um desses procedimentos.

B. ALGORITMO DE REDUÇÃO

Mediante a análise dos dados obtidos com a folha de cálculo, utilizando-se de pesquisa em matrizes com a fórmula =SOMA (SE(A4=A:A;SE(B4<>B:B;1;0);0)), que retorna a quantidade de títulos diferentes encontrados dentro da planilha (Excel) e usando como referência sua linha atual, foi constatada a falha na base de dados da plataforma Sucupira, como pode ser visto na amostra de dados obtidos no quadro 4.

Quadro 4: Amostra de dados obtidos

ISSN	Título	Área de Avaliação	Estrato	FORMULA
2257-0543	Bresil(s)	SOCIOLOGIA	B3	1
2257-0543	ARTELOGIE	INTERDISCIPLINAR	B2	4
2257-0543	Bresil(s)	HISTÓRIA	B1	1
2257-0543	Bresil(s)	CIÊNCIAS SOCIAIS APLICADAS I	B4	1
2257-0543	Bresil(s)	ANTROPOLOGIA / ARQUEOLOGIA	B3	1
2238-0167	Revista da Extensão	SAÚDE COLETIVA	C	1
2238-0167	Revista Jiop	LETRAS / LINGUÍSTICA	B5	2
2238-0167	Revista da Extensão	EDUCAÇÃO	C	1
2237-4957	Guará Linguagem e Literatura	LETRAS / LINGUÍSTICA	B5	1
2237-4957	Revista do Núcleo Onetti de Estudos Literários Latino-Americanos	LETRAS / LINGUÍSTICA	C	1
2237-3586	Vocábulo	LETRAS / LINGUÍSTICA	B4	1
2237-3586	Revista Entrepalavras	LETRAS / LINGUÍSTICA	B5	1
2236-9171	Bioenergia em Revista: Diálogos	ENGENHARIAS II	B5	1
2236-9171	Revista Diálogos Acadêmicos	ENFERMAGEM	B5	3
2236-9171	Bioenergia em Revista: Diálogos	CIÊNCIAS SOCIAIS APLICADAS I	B4	1
2236-9171	Bioenergia em Revista: Diálogos	CIÊNCIAS AGRÁRIAS I	B5	1
2236-5362	Revista da Universidade Vale do Rio Verde	INTERDISCIPLINAR	B3	2
2236-5362	CEU Arkos La Universidad Vallartense	INTERDISCIPLINAR	B4	2
2236-5362	CEU Arkos La Universidad Vallartense	ENSINO	B2	2
2236-5362	Revista da Universidade Vale do Rio Verde	ENFERMAGEM	B4	2

Com o uso da fórmula de pesquisa em matrizes em uma planilha (Excel), foi possível identificar os elementos que possuem informações inconsistentes e separar esses elementos com

mais facilidade. Embora tenha se conseguido separar os elementos na planilha, não é esse o objetivo principal (separar os elementos no momento da inserção nas tabelas), mas ajuda muito na construção do procedimento.

No banco de dados construído para o projeto, foram feitas tabelas temporárias para guardar os erros obtidos durante a inserção, para que em trabalhos posteriores os dados fossem estudados e determinar os nomes reais de tais revistas. Desse modo, as tabelas criadas são tca_imprecisos e tca_duplicados. Para tal, foi usada uma *procedure* que se baseia na lógica da Teoria dos Conjuntos Aproximativos, para reduzir a quantidade de revistas obtidas usando por critério as inconsistências encontradas durante o processo de popular o banco, tal *procedure* pode ser vista na figura 8. Para facilitar o entendimento da *procedure* foi elaborado um fluxograma, ilustrado logo a baixo pela figura 9.

Como visto no fluxograma, logo após o procedimento ter início, ele compara as informações do ISSN (chave principal) e o TITULO que estão sendo inseridas com as que já estão no banco de dados principal, caso ele constate que o ISSN esteja vinculado a um TITULO diferente do que está sendo inserido, ele classifica essa informação como imprecisa, uma vez que o ISSN é um atributo que pode facilmente ser exemplificado como uma espécie de CPF, uma identidade única para identificar os periódicos. E como não pode existir mais de uma pessoa com o mesmo CPF, também não pode existir mais de uma revista com o mesmo ISSN.

Figura 8. Fluxograma *procedure* popular_tca_reducao

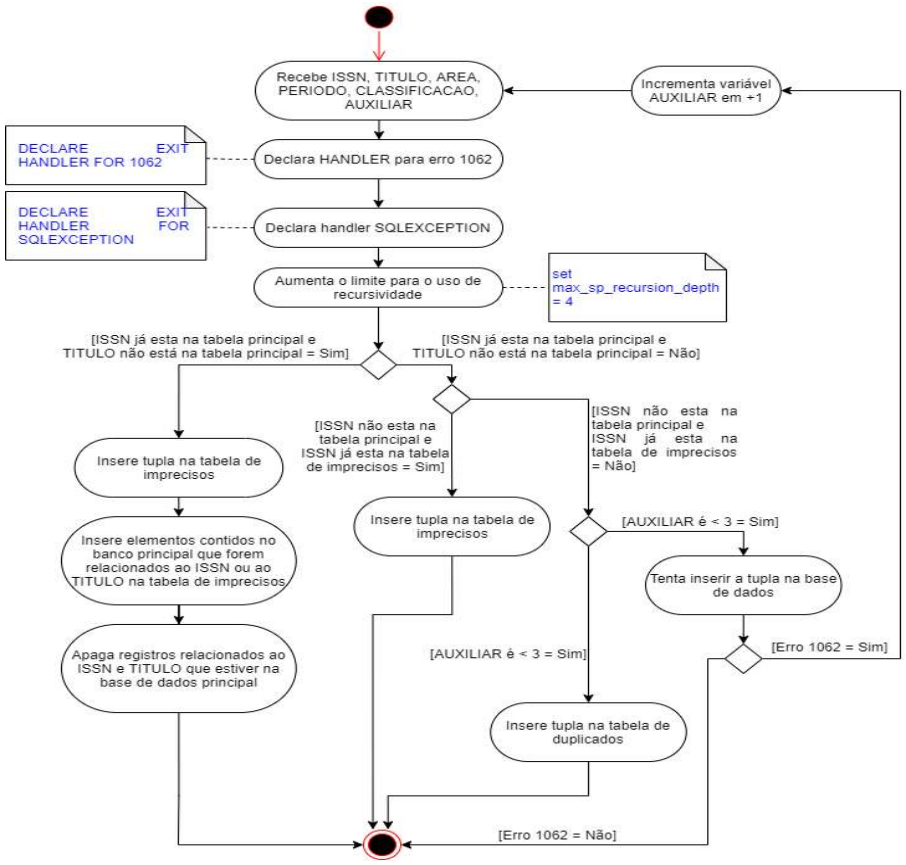
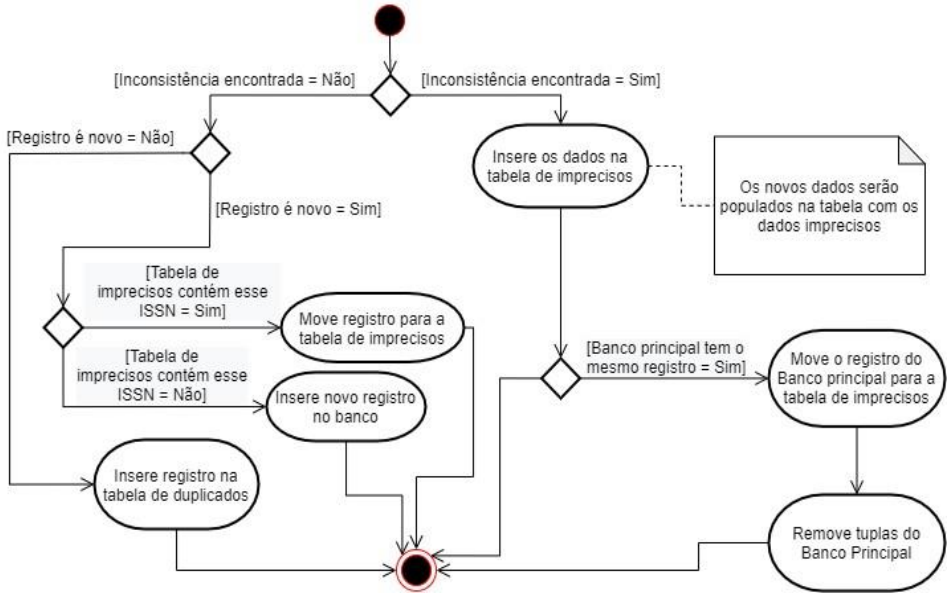


Figura 9. Fluxograma *procedure* popular_tca_reducao



Após classificar a informação como inconsistente e inserir este dado na tabela tca_imprecisos, a *procedure* chama outros dois procedimentos, popular_imprecisos e apagar_registro_por_issn, tais procedimentos são chamados para copiar as informações

relacionadas a esse ISSN impreciso, que consta na base de dados principal, para a mesma base de dados mencionada anteriormente e logo em seguida remover os elementos da base de dados. Ambos podem ser encontrados nas figuras 10 e 11 respectivamente.

Figura 10. Fluxograma *procedure* popular_impreciso

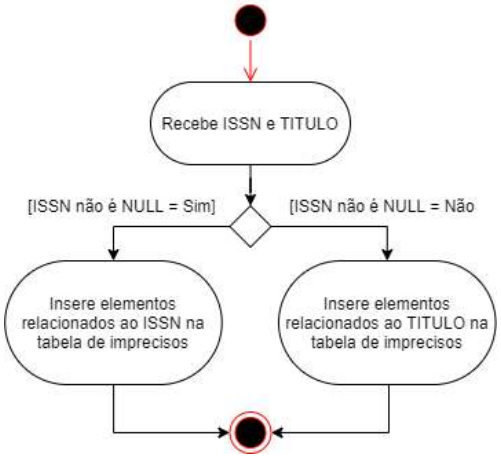
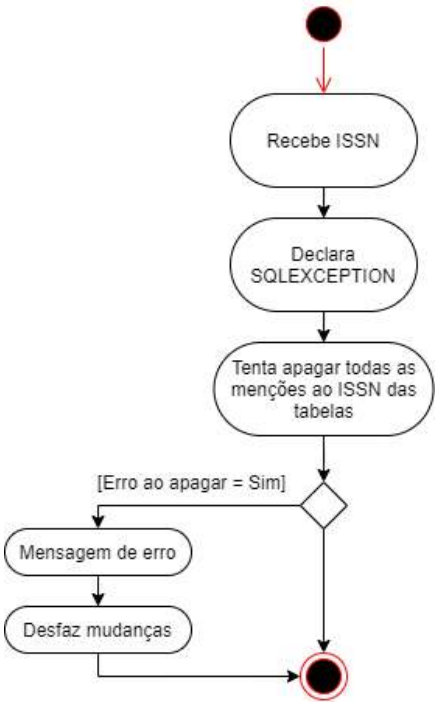


Figura 11. Fluxograma *procedure* apagar_registro_por_issn



Caso não seja identificada imprecisões aparentes, é verificado se o mesmo ISSN consta na tabela tca_imprecisos, se assim for, tal elemento é considerado impreciso e também é enviado para a mesma tabela, do contrário ele será inserido normalmente no banco de dados principal.

No meio da execução, caso venha a ocorrer o erro 1062, o procedimento identifica tal informação como “duplicada”, caso venha ocorrer, é inserido na tabela tca_duplicados.

Segue uma amostra dos dados obtidos por meio do procedimento utilizado:

Quadro 5: Amostra de dados imprecisos obtidos através do procedimento

ISSN	TITULO	AREA	PERIODO	CLASSIFICACAO
1518-5354	International Journal of Education Administration and Politic Studies	ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E TURISMO	2010 a 2013	B3
1518-5354	Resources and Environment	ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E TURISMO	2010 a 2013	B3
1518-5354	Revista Ibero-americana de Estratégia	ADMINISTRAÇÃO, CIÊNCIAS CONTÁBEIS E TURISMO	2010 a 2013	B3
1538-9472	Journal of Modern Applied Statistical Methods	CIÊNCIAS AGRÁRIAS I	2010 a 2013	B5
1538-9472	Comércio Exterior	CIÊNCIA POLÍTICA E RELAÇÕES INTERNACIONAIS	2010 a 2013	C
1516-3946	Revista do SAJU : para uma visão crítica e interdisciplinar do direito	DIREITO	2010 a 2013	C
1516-3946	Práticas jurídicas emancipatórias e arte: uma Abordagem interdisciplinar	DIREITO	2010 a 2013	C
1519-7786	Revista Síntese Direito de Família	DIREITO	2010 a 2013	C
1519-7786	Revista Iniciação Científica	DIREITO	2010 a 2013	C

C MÉTRICAS

Com os dados devidamente agrupados de acordo com a sua particularidade, sendo normal, duplicado e impreciso, foi desenvolvido algumas funções no MySQL *workbench* baseadas nos coeficientes presentes na Teoria dos Conjuntos Aproximativos (TCA), Coeficiente de Imprecisão; Coeficiente da Qualidade da Aproximação Inferior; Coeficiente da Qualidade da Aproximação Superior, assim como foi visto na seção chamada de QUALIDADE DAS APROXIMAÇÕES. Cada uma dessas funções tem sua particularidade, embora elas comumente sirvam para determinar o grau de imprecisão de um elemento.

Exclusivamente, para este trabalho foi utilizado uma tabela somente com dados imprecisos obtida por meio do algoritmo de redução mencionado na página 12, na seção ALGORITMO DE

REDUÇÃO. Entretanto, tanto no coeficiente de imprecisão quanto no coeficiente de qualidade de aproximação inferior, que medem o grau de um determinado dado ser ou não impreciso e a porcentagem de todos os objetos selecionados serem classificados como pertencentes a X, respectivamente, tiveram sua aproximação inferior fixada usando o resultado obtido em uma média simples.

Esse resultado foi obtido usando a somatória dos elementos imprecisos dos anos de 2010 e 2016 (55+5038) e a somatória do total de elementos dos mesmos anos (22048+6170) respectivamente, de modo que foi chegado no valor de 0.180487632. Isso foi feito devido a todos os dados presentes na tabela em questão serem imprecisos e, portanto, terem a confiabilidade próxima a zero.

As funções criadas seguem a lógica das métricas abordadas na Teoria dos Conjuntos Aproximativos, sendo assim, abaixo nas figuras 12, 13, 14 seguem os diagramas UML das funções criadas:

Figura 12. UML coeficiente de imprecisão

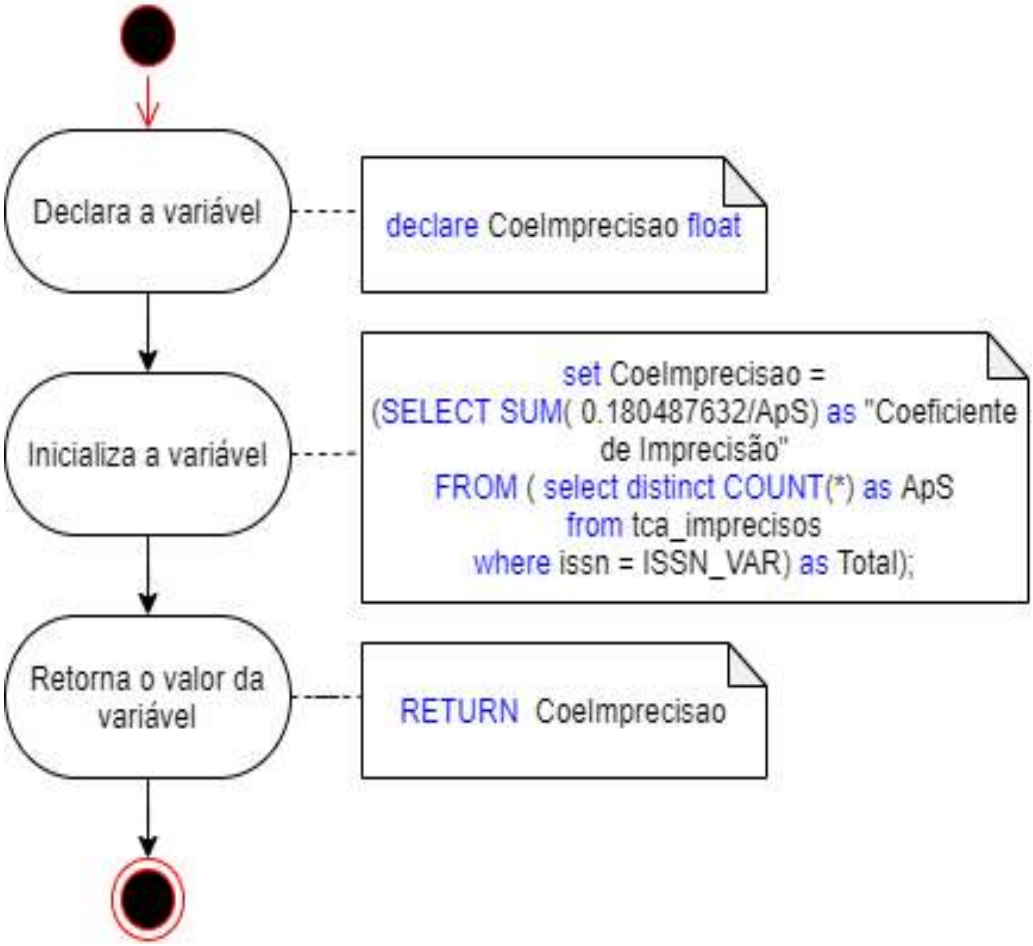


Figura 13. UML coeficiente da qualidade de aproximação inferior

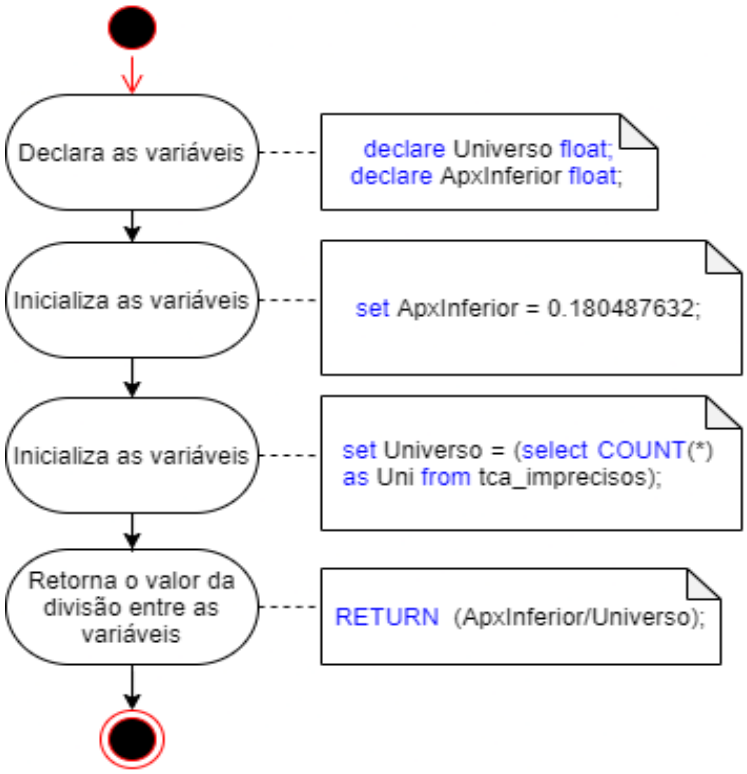
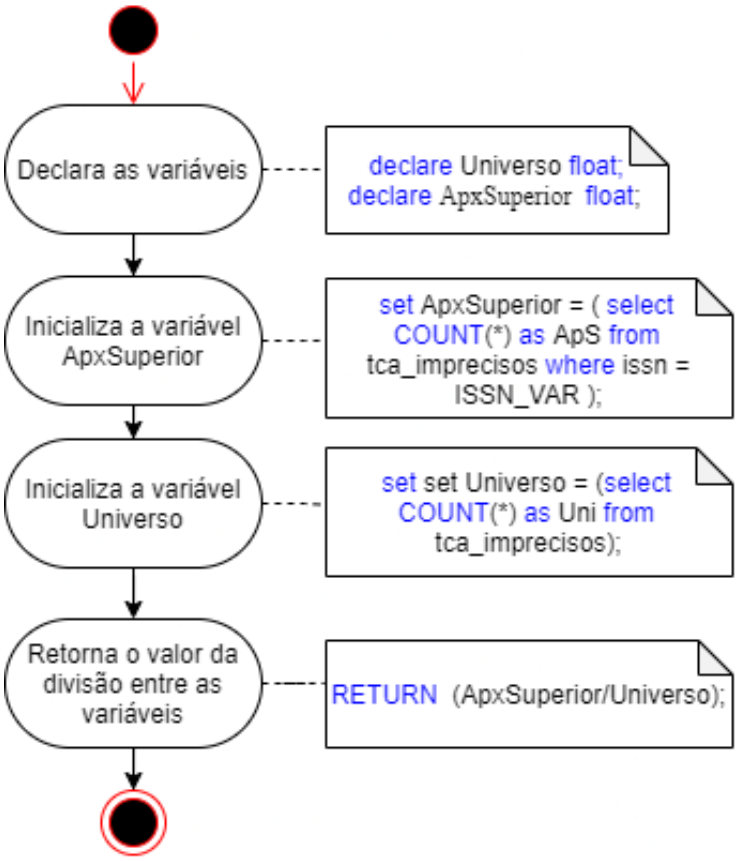


Figura 14. UML coeficiente da qualidade de aproximação superior



No quadro 6 serão exibidos os dados obtidos por meio da aplicação dos coeficientes de imprecisão nos dados obtidos por meio da *procedure* mostrada na figura 9, de que a amostra está presente no quadro 5.

Quadro 6: Amostra Dos Dados Obtidos.

ISSN	Coeficiente de Imprecisão	Coeficiente Qualidade de Aproximação Superior	Coeficiente Qualidade de Aproximação Inferior
1518-5354	0.0601625	0.00060889	0.0000366324
0000-000X	0.0601625	0.00060889	0.0000366324
9999-999X	0.022561	0.00162371	0.0000366324
1538-9472	0.0902438	0.000405927	0.0000366324
1516-3946	0.0601625	0.00060889	0.0000366324
1519-7786	0.0902438	0.000405927	0.0000366324
2176-4352	0.0300813	0.00121778	0.0000366324
2179-3565	0.0200542	0.00182667	0.0000366324
1516-3946	0.0601625	0.00060889	0.0000366324
2179-3565	0.0200542	0.00182667	0.0000366324
2236-5362	0.0451219	0.000811853	0.0000366324
0000-0000	0.0902438	0.000405927	0.0000366324
2257-0543	0.022561	0.00162371	0.0000366324

5 RESULTADOS

A Teoria dos Conjuntos Aproximativos (TCA) serviu como base para a resolução da problemática que girava em torno de imprecisões de dados. Observou-se que a utilização do procedimento de redução corroborou para uma melhor confiabilidade dos dados obtidos, por terem sido submetidos a um critério de seleção no procedimento de redução, que foi responsável por separar os elementos mais confiáveis dos elementos duvidosos.

Além disso, foi possível realizar a aplicação das métricas da TCA, por meio de funções no banco de dados MySQL *Workbench*, nas tuplas imprecisas obtidas após o processo de redução. Com os resultados das métricas, é possível identificar o grau de imprecisão dos elementos adquiridos que pode ser usado para determinar uma ordem de correção das revistas duvidosas, bem como outras finalidades.

6 CONCLUSÃO

Quando se iniciou o trabalho de pesquisa, constatou-se que os dados obtidos na plataforma Sucupira encontravam-se com alguns dados duplicados e inconsistentes, o que prejudicaria um trabalho profissional e científico em cima dessas informações. Por isso era importante estudar sobre a aplicação da Teoria dos Conjuntos Aproximativos, em incerteza de escopo de revista.

Diante disso, a pesquisa teve como objetivo geral a implementação do algoritmo da Teoria dos Conjuntos Aproximativos, ou Rough Sets em um banco de dados MySQL para resolver imprecisões de revista. Desse modo, constata-se que o objetivo geral foi atendido, já que efetivamente o trabalho conseguiu separar os atributos inconsistentes e duplicados dos saudáveis, dentro de um banco de dados MySQL.

Além disso, foi possível aplicar funções para utilizar os coeficientes de aproximações, o que foi realizado com êxito devido ter sido realizado a construção de uma função dentro do MySQL, em que tais métricas eram utilizadas.

Ademais, foi medido o grau de precisão dos termos imprecisos adquiridos, e este objetivo foi alcançado por ter sido realizado a construção de uma função que aplica as métricas da TCA nos atributos imprecisos adquiridos.

A. TRABALHOS FUTUROS:

- A criação de um procedimento ou método de busca e utilizar as tuplas imprecisas para determinar quais são seus dados reais.
- Inserir os elementos das outras tabelas (PALAVRAS_REVISTA, ARTIGOS, ARTIGOS_PALAVRAS) e aplicar coeficientes de classificação (TF-IDF, etc) para identificar o escopo de revistas.

REFERÊNCIAS

1) QUALIS. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2019. [acesso em 2019 jul. 10]. Disponível em: https://pt.wikipedia.org/w/index.php?title=Qualis&oldid=54778706.

2) Pawlak Z. Rough sets. International Journal of Computer and Information Sciences. 1982;11:341-356. [acesso em 2019 Jul. 10].

3) DATE CJ. Introdução a Sistemas de Banco de Dados. Rio de Janeiro: Campus; 1984. [acesso em 2019 Ago. 09].

4) SON SH. Replicated data management in distributed database systems. Sigmod Record. 17(4): 62-69. [acesso em 2019 Ago. 11].

5) CODD EFA. Relational model of data for large shared data banks. Communications of the ACM. 13(6):377-387. [acesso em 2019 Nov. 05].

6) MARTELLI R, FILHO OV, CABRAL AL. Modelagem e banco dedados. São Paulo: Editora Senac; 2017. [acesso em 2019 Dez. 10].

7) PESSOA ASA, SIMÕES JDS. Estudo do comportamento climático utilizando uma abordagem neuro-aproximativa. 2004. [acesso em 2020 Jan. 05]. Disponível em: http://hermes2.dpi.inpe.br:1905/col/lac.inpe.br/worcap/2004/10.06.13.09/doc/worcap_alex2004.pdf.

8) PATRÍCIO CMMM, PINTO JOP, SOUZA CC. Rough Sets – Técnica de Redução de Atributos e

Geração de Regras para Classificação de Dados. Campo Grande; 2005. [acesso em 2020 Jan. 07]. Disponível em: http://www.sbmac.org.br/eventos/cnmac/cd_xxviii_cnmac/resumos%20estendidos/cristian_patricio_ST18.pdf.

9) Suraj Z. (2004). An Introduction to Rough Set Theory and Its Applications A tutorial. [acesso em 2020 Jan. 15].

10) Cavalcanti IFA. et al. Global Climatological Features in a Simulation Using the CPTEC-COLA AGCM. Jounal of Climate. 2002;15(27):2965-2988.

11) Espaço do Conhecimento. Uma breve história da escrita. Belo Horizonte: UfMG. [acesso em 2020 Out. 21]. Disponível em: https://www.ufmg.br/espacodoconhecimento/historia-escrita/#::text=Uma%20escrita%20sistematizada%20aparece%20somente,surgem%20os%20hier%C3%B3glifos%20no%20Egito.

12) Monteiro D. A Comunicação e o tempo. Medium, 2016. [acesso em 2020 Out. 30]. Disponível em: https://medium.com/@dudamonteiro/nesse-texto-iremos-narrar-a-evolu%C3%A7%C3%A3o-da-comunica%C3%A7%C3%A3o-humana-desde-os-prim%C3%B3rdios-at%C3%A9-os-dias-atuais-91cd52510d8.

13) Sistema de gerenciamento de banco de dados. Wikipedia [acesso em 2020 Out.] Disponível em: https://pt.wikipedia.org/wiki/Sistema_de_gerenciamento_de_banco_de_dados

14) Behrouz A, Forouzan SCF. Foundations of Computer Science: From Data Manipulation to Theory of Computation. Cengage Learning Editores; 2003. ISBN 978-970-686-285-3. p. 197. [acesso em 2020 Out. 30]

15) Manual de preenchimento da Plataforma Sucupira. Sucupira, 2014. [acesso em 2020 Out. 31]. Disponível em: https://sucupira.capes.gov.br/sucupira/public/docs/manual_coleta.pdf.

16) Documentos. Github, 2017. [acesso em 2020 Nov. 04]. Disponível em: https://documentos-utilizados.github.io/.

17) Date CJ. Introdução a sistemas de bancos de dados. Rio de Janeiro: Campus; 1984.

18) Heuser CA. Projeto de banco de dados [recurso eletrônico]. 6 ed. Porto Alegre: Bookman; 2009. [acesso em 2020 Out. 10]. Disponível em: https://docplayer.com.br/87791086-Projeto-de-banco- de-dados.html.

19) Codd EF. "Relational Database: A Practical Foundation for Productivity". [acesso em 2020 Nov. 13]. Disponível em: https://dl.acm.org/doi/pdf/10.1145/358396.358400.

20) Banco de dados normalização. (Formas Normais) Diatinf. [acesso em 2020 Nov. 13]. Disponível em: http://diatinf.ifrn.edu.br/prof/lib/exe/fetch.php?media=user:1577657:26.1-bd-formas_normais.pdf.

21) KIM W. Object-Oriented Databases: Definition and Research Directions. 1990. p. 15. [acesso em 2020 Nov. 17]. Disponível em: http://130.18.86.27/faculty/warkentin/papers/8313/Kim1990.pdf.

22) SANTOS, Roger. UM ESTUDO EXPLORATÓRIO SOBRE BANCOS DE DADOS IN-MEMORY., [S. l.], p. 33, 2013. Disponível em: <https://cepein.femanet.com.br/Bdigital/arqPics/1211330014P466.pdf> Accessed: 17 nov. 2020.

23) PINTO GL. et al. Um estudo comparativo entre banco de dados relacional em disco e em memória. 2017. [acesso em 2020 Nov. 20]. Disponível em: https://repositorio.ufsc.br/handle/123456789/174203.

24) CHIKOFSKY EJ, CROSS JH. "Reverse Engineering and Design Recovery: A Taxonomy" IEEE Trans. January 1990. [acesso em 2020 Nov. 25]. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=43044.

25) Sommerville I. Engenharia de Software. 9 ed. Rio de Janeiro: Pearson Education, 2011. [acesso em 2020 Nov. 25]. Disponível em: http://www.facom.ufu.br/william/Disciplinas%202018-2/BSI-GSI030EngenhariaSoftware/Livro/engenhariaSoftwareSommerville.pdf.

26) Plataforma Sucupira. Histórico, ferramentas e tutorial para periódicos. Mettzer; 2019. [acesso em 2020 Nov. 26]. Disponível em: https://blog.mettzer.com/plataforma-sucupira/.

27) Gatto EC. Tipos de dados para uso em algoritmos. São Paulo: Embarcados; 2016. [acesso em 2020 Nov. 13]. Disponível em: https://www.embarcados.com.br/tipos-de-dados/#::text=Os%20tipos%20de%20dados%20primitivos%20s%C3%A3o%20os%20tipos%20b%C3%A1sicos%20que,%2C%20booleanos%2C%20caracteres%20e%20strings.

28) CAMILO CO, SILVA JC. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. 2009. [acesso em 2020 Nov. 28]. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf

29) Inconsistência dos dados: um dos maiores desafios frente ao eSocial. Se-nior, 2017. [acesso em 2020 Nov. 30]. Disponível em: https://www.senior.com.br/noticias/esocial-inconsistencia-dos-dados

30) INCONSISTÊNCIA. In: DICIONÁRIO da língua portuguesa. Lisboa: Priberam; 1998. . [acesso em 2020 Dez. 01]. Disponível em: https://dicionario.priberam.org/inconsist%C3%AAncia.

31) DADOS. In: DICIO, Dicionário Online de Português. Porto: 7Graus; 2020. [acesso em 2020 Dez. 01]. Disponível em: https://www.dicio.com.br/risco/.

32) PASSOS DS. Big Data: Data Science e seus contributos para o avanço no uso da Open Source Intelligence. Sistemas & Gestão. 2016;11(4):392-396.

33) Definição de banco de dados de chave-valor. [acesso em 2020 Dez. 05]. Disponível em: https://aws.amazon.com/pt/nosql/keyvalue/#::text=Um%20banco%20de%20dados%20de%20chave%2Dvalor%20%C3%A9%20um%20tipo,funciona%20como%20um%20identificador%20exclusivo.

34) O banco de dados de documentos definido. [acesso em 2020 Dez. 05]. Disponível em: https://aws.amazon.com/pt/nosql/document/.

35) Barosso I. Banco de dados orientado a colunas. [acesso em 2020 Dez. 06]. Disponível em: https://isaiasbarroso.wordpress.com/2012/06/20/banco-de-dados-orientado-a-colunas.

36) Column-Stores VS. Row-Stores: How Different Are They Really?. [acesso em 2020 Dez. 06]. Disponível em: http://www.cs.umd.edu/abadi/papers/abadi-sigmod08.pdf.

37) HEUSER CA. Projeto de banco de dados. 4 ed. Porto Alegre: Sagra- Luzzatto; 2004.

38) SASSI RJ. Aplicação dos conceitos da Teoria dos Conjuntos Aproximados no tratamento da indiscernibilidade. Exacta. 2010;8(1):89-98.

39) Sucupira. CAPES, 2017. [acesso em 2020 Fev. 24]. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.xhtml.