UNIVERSIDADE FEDERAL DO PARANÁ

RAFAEL SANFELICE CASTILHO

EXPLICAÇÃO ESTRUTURADA PARA BUSCA POR SIMILARES COMO AUXILIO PARA ANONIMIZAÇÃO DE DADOS

CURITIBA PR

2021

RAFAEL SANFELICE CASTILHO

EXPLICAÇÃO ESTRUTURADA PARA BUSCA POR SIMILARES COMO AUXILIO PARA ANONIMIZAÇÃO DE DADOS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Fabiano Silva.

CURITIBA PR

2021

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

C352e Castilho, Rafael Sanfelice
Explicação estruturada para busca por similares como auxílio para anonimização de dados [recurso eletrônico] / Rafael Sanfelice Castilho – Curitiba, 2021.

Dissertação - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-graduação em Informática.

Orientador: Fabiano Silva.

1. Proteção de dados. 2. Anonimização (Informática). I. Universidade Federal do Paraná. II. Silva, Fabiano. III. Título.

CDD: 005.85

Bibliotecária: Roseny Rivelini Morciani CRB-9/1585

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIENCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA - 40001016034P5

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **RAFAEL SANFELICE CASTILHO** intitulada: **EXPLICAÇÃO ESTRUTURADA PARA BUSCA POR SIMILARES COMO AUXILIO PARA ANONIMIZAÇÃO DE DADOS**, sob orientação do Prof. Dr. FABIANO SILVA, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 17 de Dezembro de 2021.

Assinatura Eletrônica
06/01/2022 17:54:23.0
FABIANO SILVA
Presidente da Banca Examinadora

Assinatura Eletrônica
07/01/2022 08:11:29.0
LUIZ CELSO GOMES JUNIOR
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

Assinatura Eletrônica
06/01/2022 17:19:28.0
MARCOS DIDONET DEL FABRO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

A minha família, que sempre me incentivou e me deu apoio todo esse tempo

AGRADECIMENTOS

Agradeço à minha família por ter me incentivado a perseguir meus estudos e interesses, e por ter me dado apoio ao longo de toda a vida de estudante. Principalmente durante o desenvolvimento do mestrado, em que tive muitos momentos de stress em que fiquei extremamente rabugento.

Um agradecimento especial a minha mãe, que fez a revisão do texto mesmo estando muito ocupada, sem ficar louca com meus diversos erros de português e estilo falado de escrever textos.

Ao meu orientador, o professor Dr. Fabiano Silva, que me norteou quando estava perdido sem saber como prosseguir, e teve a paciência de me orientar mesmo enquanto trabalhava como chefe de departamento no período caótico de pandemia em que fiz a maior parte do mestrado.

À Maria Helena Fransiscato por disponibilizar o código dela que fazia parte do processo de busca por similares e estar disposta a me ajudar a entender todas as partes que me deixaram com dúvidas na parte do trabalho dela que coincidia com o meu.

Aos professores doutores Marcos Didonet del Fabro e Luis Carlos Erpen de Bona p or terem se disponibilizado a participar de discussões sobre o rumo da pesquisa no começo, quando não tinha uma direção a ser seguida, e me ajudar a encontrar o caminho a seguir.

Ao C3SL por me proporcionar um bom ambiente de trabalho em que pude trabalhar com diversas tecnologias e interagir com diversas pessoas com quem pude conversar a respeito de como estava me sentindo com o mestrado e trocar técnicas para lidar com stress. Além de prover o ambiente de trabalho para que eu realizasse meus experimentos.

RESUMO

Vivemos em uma época em que, cada vez mais, dados de qualquer pessoa são públicos, ou podem tornar-se públicos, seja em função de iniciativas de transparência em dados abertos, seja devido a publicações sem consentimento da pessoa a quem os dados pertencem. Isso apresenta um problema para a privacidade de indivíduos, pois aumenta as chances de alguém com más intenções desfazer a anonimização de dados presentes em uma publicação que contenha dados sensíveis. Diversos trabalhos em anonimização lidam com o problema de manter a privacidade de indivíduos cujos dados sensíveis estejam presentes em publicações. Ao mesmo tempo, diversos trabalhos em integração de dados e buscas por similares lidam com prover uma visão unificada sobre dados publicados e encontrar relações entre eles. Apesar disso, ao longo do desenvolvimento desta dissertação não foi encontrado nenhum trabalho que explicitasse a relação entre anonimização, integração e busca por similares ou que utilize esta relação de alguma forma. Neste trabalho apresentamos nossa proposta para uma forma de estruturar explicações do resultado de uma busca por similares em dados abertos em estruturas de dados hierárquicas para fins de auxiliar o processo de anonimização.

Palavras-chave: Similares. Anonimização. Explicação.

ABSTRACT

We live in times where more and more personal data is public, or can be made public, either by an open data initiative, or by being published without the consent of the individual to whom the data originates from. This is a problem for personal privacy, since it increases the likelihood of someone with malicious intent undoing the anonymization on the published data that contains sensitive information. Many papers on anonimity deal with the problem of protecting the privacy of an individual whose sensitive data is present in a data publication. At the same time, many papers and open data integration and open data search deal with providing an overarching view on data publications and finding relation between said data. In spite of that, during the elaboration of this dissertation no paper was found that explicitly talked about the relationshiip between anonization, open data interation and open data search, or at least that used this relationship in any manner. In this dissertation we present our proposal for a method of structuring the explanation about the result of an open data search in a hierarchical data structure, to help the data anonymization process.

Keywords: Similarity. Anonymization. Explanation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE LISTAGENS

SUMÁRIO

1 INTRODUÇÃO

O objetivo deste trabalho é expor a relação que existe entre as áreas de anonimização e busca por similares, como elas possuem um relacionamento adverso, mas que se utilizada corretamente esta relação pode trazer benefícios para a anonimização de dados. Parte deste processo é facilitado se a busca por similares efetuada for explicada, portanto de forma secundaria este trabalho lida com a construção de uma explicação para o processo de busca por similares e estuda como fazer uma boa explicação.

Anonimização de dados é a prática utilizada para salvaguardar a privacidade de um indivíduo cujas informações pessoais estejam presentes em uma base de dados abertos. Para que isso seja possível,uma entidade (empresa, órgão do governo, indivíduo, etc) manipula esses dados antes de publicá-los, de forma que não seja mais possível conectar os dados presentes na base de dados publicada com o indivíduo a que se referem. Essa manipulação pode consistir tanto em uma edição dos dados para torná-los mais genéricos, quanto na completa remoção deles. Essa manipulação é feita a partir da classificação dos atributos – características dos indivíduos –, que podem ser separados de dois modos diferentes (Merener, 2012; Murthy et al., 2019; Ozalp et al., 2016).

O primeiro desses dois modos é a divisão entre atributos sensíveis e não-sensíveis. Atributos sensíveis, como o nome evidencia, são aqueles que contêm informações que não deveriam ser conectadas a indivíduos, mas cuja presença é relevante na publicação de dados; por exemplo, uma informação sobre a condição de saúde. Por outro lado, atributos não-sensíveis são aqueles que, em princípio, não revelam nada que possa ser considerado extremamente confidencial a respeito de um indivíduo, como, idade ou gênero.

O segundo modo separa os atributos de acordo com sua capacidade de identificar um indivíduo. Identificadores explícitos, como o nome aponta, identificam de maneira explícita um indivíduo (por exemplo um CPF) espera-se que, no processo de anonimização, todos os identificadores explícitos sejam removidos. Quase-identificadores (Qis), por outro lado, são atributos que podem ser usados para identificar um indivíduo quando combinados com atributos diferentes, por exemplo, data de nascimento e CEP, por si só, não identificam um indivíduo, mas combinados podem ser usados para tal tarefa. Determinar se um atributo é ou não um quase-identificador é uma tarefa difícil, pois depende de conhecimento sobre os atributos publicados. Por fim, em uma categoria sem nome definido, está o resto dos atributos, qualquer atributo neste conjunto é um potencial quase-identificador, basta encontrar uma combinação de atributos que possa ser utilizada para identificar indivíduos, mesmo que seja necessário combinar atributos de diversas tabelas.

Desse modo, manter a utilidade de dados anonimizados e, ao mesmo tempo, preservar a privacidade dos indivíduos aos quais esses dados pertencem é o principal desafio no campo da anonimização. Grande parte deste problema vem do fato que não é possível saber a que informações um atacante (indivíduo ou grupo malicioso que tem como objetivo desfazer a anonimização executada e obter as informações sensíveis sobre indivíduos presentes na tabela) tem acesso, então as entidades que publicam dados anonimizados costumam assumir o pior caso – situação em que o atacante deteria todas as informações –, consequentemente maximizando a anonimização e minimizando a utilidade dos dados.

Nossa hipótese é que a anonimização tem uma relação dual com a busca por similares em dados abertos, uma área de pesquisa que se ramificou a partir da integração de dados abertos.

Integração de dados abertos é uma área de pesquisa cujo objetivo é prover uma visão unificada dos dados em um banco. Uma das formas mais comuns de fazer isto é gerar mapeamentos entre diferentes tabelas, presentes em repositórios de dados abertos. Os mapeamentos gerados podem ser entre tabelas do mesmo repositório ou de repositórios diferentes (Berlin e Motro, 2002).

Essas tabelas contêm dados dispostos em linhas e colunas. Cada linha (ou elemento) se refere a um indivíduo (ou entidade) representado pelos dados. Cada coluna representa uma característica (ou atributo) dos dados, como nome ou idade, por exemplo.

Um mapeamento é feito a partir de diferentes atributos de um par (ou conjunto) de tabelas para determinar se os dados representados contêm a mesma informação (por exemplo, se duas tabelas contêm o atributo "nome", mesmo que o nome do atributo seja diferente). Um mapeamento indica, para um determinado par (ou conjunto) de tabelas, as colunas que as tabelas têm em comum.

No contexto de integração de dados, o objetivo de gerar mapeamentos entre tabelas é possibilitar a integração dos dados entre elas. Um atributo pode ser mapeado para um único atributo em outra tabela, ou para diversos atributos em outra tabela, esta cardinalidade depende da estratégia de mapeamento sendo utilizada.

Similaridade é o nome dado a esse fenômeno das tabelas (ou colunas) conterem o mesmo tipo de informação – por exemplo nome, ou CEP. A diferença é que na integração de dados o objetivo é obter um mapeamento que indique que os dados definitivamente representam a mesma informação, enquanto em busca por similares basta obter um conjunto de candidatos similares para análise futura (Nargesian et al., 2018).

A busca por similares é uma parte do processo de integração de dados que passou a receber um foco separado da integração por ter utilidades além de seu propósito em integração.

Existem diversas maneiras de se fazer uma busca por similares, porém, para esta dissertação e nos trabalhos estudados a similaridade é encontrada com o uso de tabelas de bases de dados. Neste contexto uma busca por similares consiste em percorrer um repositório de dados e encontrar tabelas que sejam similares com relação a uma determinada tabela para alguma métrica (ou métricas) de comparação.

Dependendo do objetivo de quem faz a busca, ela pode ser feita com qualquer subconjunto de colunas da tabela de base para fazer as comparações. Devido à grande quantidade de tabelas em repositórios de dados abertos, não é desejável que o resultado da busca seja um ranqueamento de todas as similaridades calculadas no repositório, mas sim um conjunto limitado com as k tabelas mais similares, com k sendo um número arbitrariamente escolhido por quem realiza a busca para limitar o tamanho do conjunto de tabelas que se deseja encontrar.

Para calcular a similaridade entre duas tabelas, são combinadas as similaridades entre seus atributos. Existem diversas métricas para comparar os dados dos atributos, com diferentes desempenhos, dependendo do tipo dos dados que estão sendo comparados (texto, números inteiros, números de ponto flutuante, etc).

A forma como fizemos o cálculo de similaridade tem a restrição de que a geração de mapeamentos em que um atributo de uma tabela só deveria ser considerado similar a, no máximo, um atributo da outra tabela. A determinação de quais são os pares de atributos similares e a combinação das similaridades entre atributos, a fim de definir um valor para a similaridade entre tabelas, podem ser feitas de diversas formas, algumas das quais serão apresentadas na seção 2.1.

A relação dual entre anonimização e busca por similares vêm do fato da anonimização remover informações das tabelas publicadas, afetando a qualidade da busca, já que tabelas anonimizadas oferecem menos informações para determinar similaridades.

Por outro lado, a busca por similares apresenta riscos para a anonimização, já que uma busca por similares pode ser utilizada para encontrar um conjunto de tabelas similares a uma tabela anonimizada e, a partir disso, obter um conjunto de QIs, que podem ser utilizados para quebrar a anonimização.

Outro desafio, secundário, mas também complicado, é como apresentar o resultado da busca por similares de forma a melhor auxiliar na análise de riscos para a anonimização. Explicar o processo de busca é a melhor forma de fazer isso, porém, definir a qualidade de uma explicação é difícil, já que esta varia muito dependendo do contexto em que a explicação é provida.

Diante dessa perspectiva, nossa motivação para este estudo foi melhorar a análise de risco feita antes da anonimização, para que as entidades que anonimizam dados tenham uma noção mais razoável a que dados um possível atacante tem acesso.

Para prever quais dados são mais vulneráveis ao acesso de um atacante, fazemos uma busca por similares e utilizamos as tabelas similares à tabela a ser anonimizada como base para a análise de riscos da anonimização. Para lidar com os desafios de explicar o resultado da busca por similares, estudamos como fazer uma boa explicação que seja mais flexível para mudanças de contexto e grupo alvo.

Nossa proposta é apresentar um modo de estruturar explicações de forma que sejam facilmente manipuladas para se adaptarem a diversos contextos. Essa estruturação permite ao provedor da explicação maximizar o uso de fatores que melhoram a qualidade dela, enquanto minimiza os fatores que a pioram. Além disso, propomos o uso desse modelo em uma busca por similares para fazer uma análise de riscos mais informada durante o processo de anonimização de dados, facilitando a utilização da busca por similares como parte do processo de anonimização de dados.

As contribuições deste trabalho para a área do conhecimento em questão se dão no âmbito da busca por similares e na geração de explicações. Assim, elaboramos um modelo de utilização de busca por similares para melhorar a análise de riscos efetuada durante o processo de anonimização de dados. Além disso, construímos um modo de gerar explicações estruturadas, de maneira que uma mesma explicação possa ser apresentada de forma flexível a diversos grupos alvos diferentes.

Esta dissertação se organiza em 5 capítulos. No capítulo 2 apresentamos os trabalhos usados para fundamentar esta dissertação. No capítulo 3 apresentamos com maiores detalhes nossa proposta. No capitulo 4 descrevemos os experimentos realizados e os resultados obtidos. Por fim, no capítulo 5, apresentamos as conclusões a que chegamos ao final da pesquisa.

2 TRABALHOS RELACIONADOS

Neste capítulo, apresentamos os estudos que embasaram esta dissertação. Inicialmente, na seção 2.1 referenciamos estudos da área de busca por similares em dados abertos, que apresentam a área, o modo como ela se ramificou a partir de estudos sobre integração, e técnicas utilizadas para medir similaridade.

Na seção 2.2 sintetizamos os trabalhos referentes à anonimização de dados, às técnicas utilizadas nesse processo, aos problemas que existem na área, em particular ataques de inferência, e à dificuldade de balancear privacidade e utilidade.

Por fim, na seção 2.3 discutimos os estudos referentes à explicação de sistemas – em particular sistemas com IA e aprendizado de máquinas –, que esclarecem a importância de se ter explicações e algumas das dificuldades encontradas em trabalhos na área, como a falta de uma nomenclatura padrão ou a dificuldade em definir métricas de qualidade para explicações.

2.1 BUSCA POR SIMILARES EM DADOS ABERTOS

Integração de dados é uma área de pesquisa cujo objetivo é prover uma visão unificada dos dados em um banco. Uma das formas mais comuns de fazer isto é gerar mapeamentos entre diferentes tabelas, presentes em bancos de dados. Os mapeamentos gerados podem ser entre tabelas do mesmo repositório ou de repositórios diferentes.O objetivo de gerar esses mapeamentos entre dados é agregação de informações.

A área de integração de dados é histórica e diversos trabalhos foram publicados para trabalhar com ela, porém neste trabalho a integração de dados é usada apenas para contextualizar a área de busca por similares.

A busca por similares é uma parte do processo de integração de dados que passou a receber um foco separado da integração por ter utilidades além de seu propósito em integração. Uma busca por similares consiste em percorrer um repositório de dados e encontrar tabelas que sejam similares com relação a uma determinada tabela para alguma métrica (ou métricas) de comparação.

Os trabalhos Nargesian et al. (2018); Miller (2018); Zhu et al. (2019); Bogatu et al. (2020) apresentam diferentes maneiras de efetuar a busca por similares, tanto em termos de formas de comparação, quanto em termos de seleção de tabelas para comparar.

Ao nosso entendimento Nargesian et al. (2018) e Miller (2018) são trabalhos publicados pelo mesmo grupo sobre o mesmo trabalho, com focos diferentes.

O texto de Nargesian et al. (2018) foca na discussão de técnicas de comparação de atributos e tabelas, avaliando o desempenho de quatro formas de comparar atributos.

Por outro lado o texto de Miller (2018) discute a separação da busca por similares da área de integração e por que é necessária, além de apresentar uma estrutura de indexação que pode ser utilizada para filtrar candidatos de comparação.

Os trabalhos Zhu et al. (2019); Bogatu et al. (2020) evoluem a proposta de Nargesian et al. (2018) e Miller (2018). O texto de Zhu et al. (2019) inclui a partcipação de Narguesian, e foca em utilizar a estrutura apresentada por Miller (2018) em conjunto com um par de filtros para otimizar a seleção de candidatos para comparação, com o objetivo de reduzir o numero de tabelas que são comparadas com a tabela base da busca.

O Trabalho de Bogatu et al. (2020) segue a metodologia de Nargesian et al. (2018), porém com diferentes técnicas de comparação, tanto de similaridade entre atributos, quanto de

similaridade entre tabelas. Além das técnicas de comparação, outra diferença trazida por Bogatu et al. (2020) é a ideia de usar similaridade para medir relacionamento entre os dados.

Nenhum dos trabalhos apresentados discute como lidar com o problema de duas tabelas sendo comparadas possam ter uma diferença grande no numero de elementos ou atributos. Essa diferença – dependendo de como as comparações são feitas – podem levar a um viés grande nos resultados.

2.2 ANONIMIZAÇÃO

Anonimização de dados é a prática utilizada para salvaguardar a privacidade de um indivíduo cujas informações pessoais estejam presentes em uma base de dados abertos. Para que isso seja possível, uma entidade (empresa, órgão do governo, indivíduo, etc) manipula esses dados antes de publicá-los, esperando que não seja mais possível conectar os dados presentes na base de dados publicada com o indivíduo a que se referem. Essa manipulação pode consistir tanto em uma edição dos dados, quanto na completa remoção deles.

Os trabalhos de Sweeney (2002); Machanavajjhala et al. (2007); Li et al. (2007); Xiao e Tao (2007); Brickell e Shmatikov (2008); Murthy et al. (2019); Ozalp et al. (2016) apresentam diversas técnicas para manipular e organizar os dados de forma a evitar desanonimização dos dados. Separadamente, o trabalho de Merener (2012) discute um dos ataques feitos para desanonimizar dados já publicados.

O texto de Murthy et al. (2019) apresenta cinco técnicas básicas de manipulação de dados para efetuar a anonimização, e avalia seus desempenhos. Concluindo que não ha uma técnica perfeita de anonimização, e que, para cada anonimização efetuada deve ser feita uma avaliação para decidir qual a técnica mais adequada a ser utilizada para os dados em questão.

Já os trabalhos de Sweeney (2002); Machanavajjhala et al. (2007); Li et al. (2007); Xiao e Tao (2007); Brickell e Shmatikov (2008); Ozalp et al. (2016) focam em técnicas de organização dos dados após a aplicação de técnicas de manipulação terem anonimizado os dados. Estas técnicas todas se baseiam em verificar se os dados anonimizados atendem a uma condição de organização, caso não atendam é necessário aplicar as técnicas bases para anonimizar mais os dados até que atendam a condição necessária. Com exceção de Ozalp et al. (2016) todos os trabalhos lidam com anonimização no contexto de tabelas de bancos de dados, Ozalp et al. (2016) por outro lado foca em anonimização do que chama de registros hierárquicos.

Para Sweeney (2002) a condição é a k-anonimidade, que requer que para o conjunto de atributos identificados como quase-identificadores seus valores apareçam k vezes, como pode ser observado na tabela 2.1.

Por sua vez Machanavajjhala et al. (2007) critica Sweeney (2002), afirmando que não basta que os quase-identificadores estejam organizados em conjuntos de k instancias, mas que também é necessário que para estes conjuntos os dados sensíveis estejam representados por pelo menos l valores distintos. Esta condição é chamada de l-diversidade e é uma forma mais restrita de k-anonimização, um exemplo pode ser encontrado na tabela 2.2.

De forma similar o trabalho de Li et al. (2007) aponta falhas em Sweeney (2002); Machanavajjhala et al. (2007) e propõe a condição de t-proximidade para evitar a desanonimização dos dados. Para Li et al. (2007) a solução é que a distribuição de valores distintos para dados sensíveis nos conjuntos k-anônimos deve ser próxima a sua distribuição no conjunto total de dados dentro de um limiar t. Um exemplo de tabela t-próxima pode ser observado em 2.3.

Para Xiao e Tao (2007) os trabalhos de Sweeney (2002); Machanavajjhala et al. (2007); Li et al. (2007) possuem a mesma falha, eles não levam em consideração a possibilidade dos dados serem republicados em uma data futura. Xiao e Tao (2007) propõem então a condição de

Race	Birth	Gender	ZIP	Problem
Black	1965	m	0214*	short breath
Black	1965	m	0214*	chest pain
Black	1965	f	0213*	hypertension
Black	1965	f	0213*	hypertension
Black	1964	f	0213*	obesity
Black	1964	f	0213*	chest pain
White	1964	m	0213*	chest pain
White	1964	m	0213*	obesity
White	1964	m	0213*	short breath
White	1967	m	0213*	short breath
White	1967	m	0213*	chest pain

Tabela 2.1: Tabela 2-anonima para Race, Birth, Gender e ZIP retirada de (Sweeney, 2002)

Non-Sensitive			Sensitive
Zip Code	Age	Nationality	Condition
1305*	<= 40	*	Heart Disease
1305*	<= 40	*	Viral Infection
1305*	<= 40	*	Cancer
1305*	<= 40	*	Cancer
1485*	>40	*	Cancer
1485*	>40	*	Heart Disease
1485*	>40	*	Viral Infection
1485*	>40	*	Viral Infection
1306*	<= 40	*	Heart Disease
1306*	<= 40	*	Viral Infection
1306*	<= 40	*	Cancer
1306*	<= 40	*	Cancer

Tabela 2.2: Tabela 3-diversa retirada de (Machanavajjhala et al., 2007)

	ZIP Code	Age	Salary	Disease
1	4767*	<= 40	3K	gastric ulcer
3	4767*	<= 40	5K	stomach cancer
8	4767*	<= 40	9K	pneumonia
4	4790*	>= 40	6K	gastritis
5	4790*	>= 40	11K	flu
6	4790*	>= 40	8K	bronchitis
2	4760*	<= 40	4K	gastritis
7	4760*	<= 40	7K	bronchitis
9	4760*	<= 40	10K	stomach cancer

Tabela 2.3: Tabela 0.167-próxima para Salary e 0.278-próxima para Disease retirada de (Li et al., 2007)

m-invariância, que exige que exista um grau m de invariância nos valores que dados sensíveis recebem nos conjuntos k-anônimos em publicações distintas dos dados.

Por fim, o trabalho de Brickell e Shmatikov (2008) faz uma critica aos trabalhos de Sweeney (2002); Machanavajjhala et al. (2007); Li et al. (2007); Xiao e Tao (2007), além de

criticar a forma como trabalhos em anonimização medem a perda de utilidade nos dados causados pela anonimização. A forma de organização proposta por Brickell e Shmatikov (2008) é similar a t-proximidade, mas ao invés de exigir que a distribuição de valores para dados sensíveis no conjunto k-anonimo seja próxima a distribuição no conjunto total de dados, exige que a distribuição seja similar.

Quanto as criticas na medição de perda de utilidade, afirma que o que é medido é o numero de operações feitas nos dados, e não a perda de utilidade. Concluindo com a afirmação de que a utilidade dos dados é muito relativa a como são utilizados, e portanto não é possível definir uma utilidade estática para os dados.

Como dito no inicio da seção, o trabalho de Ozalp et al. (2016), diferente dos outros trabalhos apresentados, não lida com tabelas de bancos de dados e sim com o que chama de registros hierárquicos. De forma simples, registros hierárquicos são arvores de dados, com as folhas contendo os dados e os nós sendo atributos. Ozalp et al. (2016) propõe formas de aplicar as técnicas de anonimização de tabelas no contexto de registros hierárquicos, definindo também k-anonimidade e l-diversidade no contexto de registros hierárquicos.

Os trabalhos apresentados até aqui tratam todos de diferentes formas de manipular dados que serão publicados para reduzir a chance de desanonimização. O próximo trabalho que discutiremos, Merener (2012) não lida com a manipulação dos dados, mas sim com um dos ataques que são realizados em dados já anonimizados e publicados.

O ataque apresentado é o chamado ataque de ligação (*Linkage attack*). De forma simples, um ataque de ligação utiliza um conjunto de quase-identificadores para desanonimizar os dados, o que torna difícil de preveni-lo é que um ataque de ligação combina atributos de diversas tabelas para construir o conjunto de quase-identificadores.

O trabalho de Merener (2012) discorre sobre, o que são ataques de ligação, como funcionam, como precisam de uma quantidade baixa de atributos para terem sucesso, e como podem ser combatidos.

2.3 EXPLICAÇÃO

Uma explicação é – de forma bem abrangente – o processo de trazer entendimento a respeito de algo para um determinado grupo alvo. A qualidade de uma explicação esta atrelada a capacidade do grupo alvo de entender a explicação provida, e diversos fatores podem afetá-la.

Nesta sessão apresentamos os trabalhos estudados durante o desenvolvimento desta dissertação. Ao longo da pesquisa foi encontrado apenas um trabalho que menciona explicações no contexto de integração de dados, portanto, o restante dos trabalhos apresentados é referente à explicações em IA de maneira mais geral.

Nesta seção serão apresentados trabalhos com diferentes objetivos, na subseção 2.3.1 são apresentados os trabalhos cujo foco esta em caracterizar diferentes aspectos sobre a construção de explicações, enquanto na subseção 2.3.2 estão os trabalhos referentes a métricas de qualidade para explicações.

2.3.1 Caracterização

Os trabalhos de Wang et al. (2018); Došilović et al. (2018); Bohlender e Köhl (2019) tratam de caracterizar a explicação em si, tentando definir o que é uma explicação, que características possui e como podem ser classificadas.

Por outro lado o trabalho de Hoffman et al. (2018) foca no aspecto de qualidade de uma explicação, o que afeta a qualidade de uma explicação e o que deve ser levado em consideração para medir a qualidade de uma explicação.

O trabalho de Wang et al. (2018) foi o unico trabalho encontrado que menciona a necessidade de prover explicações para o processo de integração de dados, mesmo assim o foco principal do texto esta em classificar sistemas com relação a sua capacidade de prover explicações.

Os autores primeiro separam explicações em causais – que explicam como um resultado foi obtido – e não-causais – que explicam qual foi o resultado obtido. Em seguida separam sistemas de acordo com como provém explicações, separando-os em sistemas explicadores, sistemas explicáveis, sistemas ilustrativos e sistemas não explicáveis.

Por fim os autores definem dois eixos que dever ser levados em consideração para classificar explicações. O primeiro é o eixo de cobertura, que se refere ao detalhamento de uma determinada explicação, quanto mais detalhada a explicação provida, maior a cobertura. O segundo eixo é o eixo de legibilidade, quanto menos conhecimento prévio necessário para entender a explicação maior é o grau de legibilidade. Quanto maior a cobertura menor a legibilidade e vice-versa.

A pesquisa de Došilović et al. (2018) analisa diversos trabalhos a respeito da geração de explicação, e aponta problemas na falta de um consenso existente para terminologia nos trabalhos publicados. O trabalho propõe algumas formas de lidar com esta falta de consenso.

O trabalho de Bohlender e Köhl (2019) aponta a necessidade de se ter uma caracterização comum para explicações, de forma que seja possível classificar, comparar e avaliar diferentes abordagens para um sistema. A caracterização proposta pelos autores envolve separa a explicação em seus componentes básicos, e em explicitar noções que outros trabalhos utilizam de forma menos explicita.

Os autores sugerem, por fim, como diferentes áreas podem trabalhar com explicações. As sugestões vão desde como melhorar a representação dos dados (interação humano-computador, psicologia, etc) e como definir requisitos para uma explicação (engenharia de requisitos), até a geração da explicação (IA e aprendizado de máquina).

O texto de Hoffman et al. (2018) discute o que caracteriza uma boa explicação, incluindo fatores que melhoram ou pioram a qualidade de uma explicação. Os autores se baseiam em diverso trabalhos em psicologia para entender como humanos processam explicações que recebem, e fatores que facilitam a recepção de uma explicação.

Um apontamento importante feito pelos autores é que a qualidade de uma explicação é muito dependente do grupo alvo para quem a explicação é provida, dessa forma não é possível medir a qualidade de uma explicação, é possível apenas medir a qualidade de uma explicação para um determinado grupo alvo, dessa forma, não é desejável que seja feita uma métrica de qualidade descontextualizada.

Todos os trabalhos apresentados reforçam a necessidade de programas e sistemas proverem explicações, principalmente por que o uso de sistemas esta cada vez mais entrelaçado com o cotidiano humano, e é necessário que pessoas possam confiar nesses sistemas.

2.3.2 Métricas de qualidade

O trabalho de Holzinger et al. (2020) definem uma escala para medir a qualidade de uma explicação provida com a utilização de um questionário que deve ser preenchido por quem recebeu a explicação. O questionário apresenta uma serie de afirmações em que uma pessoa deve quanto concorda com as afirmações, desde “discordo fortemente” até “concordo fortemente”, a partir das respostas é calculado uma pontuação entre 0 e 100 para a qualidade da explicação.

Já o trabalho de Rosenfeld (2021) critica o uso de questionários para medir a qualidade de uma explicação, citando que não capturam corretamente a dinamica entre performance e explicação, além de serem suscetíveis a viés de confirmação.

Os autores propõem quatro métricas para avaliar a qualidade de uma explicação, apesar de não estar explicito no texto, como todas as métricas propostas são implementadas como um peso na performance – do que esta sendo medido a performance não é definido – entendemos que as métricas propostas são para ferramentas de aprendizado de máquina.

Dentre as métricas propostas, a primeira mensura a diferença de performance entre um modelo de aprendizado caixa preta e o melhor modelo transparente que estiver disponível. A segunda, calcula o tamanho da explicação gerada, comparando o número de regras estabelecidas com um número mínimo de regras aceitáveis. A terceira métrica mede quantas regras o agente recebe de entrada, comparando-as com um número mínimo aceitável de regras de entrada. A quarta, e última, métrica pondera a estabilidade da explicação quando é feito um *bootstrapping* da entrada.

2.4 CONSIDERAÇÕES

Os trabalhos apresentados neste capítulo mostram os desenvolvimentos encontrados nas áreas de busca por similares, anonimização e explicação de resultados. A partir dessas leituras, percebemos as relações que existem entre essas áreas, além do evidente relacionamento entre integração de dados e busca por similares em dados abertos.

A primeira relação percebida foi entre busca por similares e anonimização. A percepção inicial foi de que o relacionamento é antagônico, pois a anonimização de dados, por remover informações dos dados, dificulta o processo de busca. Ao mesmo tempo, a busca por similares, por encontrar de forma eficiente dados similares que podem ser relacionados entre si, traz riscos para a anonimização, pois os dados encontrados podem incluir quase-identificadores ainda desconhecidos durante a anonimização dos dados.

Apesar desse antagonismo, é possível mitigar essa diferença. Se a busca for utilizada antes da anonimização e de preferência em múltiplos repositórios, a entidade responsável pela anonimização é capaz de fazer uma análise de riscos muito mais detalhada, com mais segurança de quais dados são quase-identificadores; com essa certeza, é possível melhorar o processo de anonimização, reduzindo o risco de desanonimização.

O benefício secundário de usar o resultado de uma busca por similares para a análise de riscos da anonimização é a possibilidade de realizar a busca novamente após a anonimização dos dados e comparar os resultados obtidos e, com isso, fazer uma análise de qualidade da anonimização. A intuição é que uma anonimização ruim teria resultados similares para as tabelas que contivessem quase-identificadores. Ao mesmo tempo, a possibilidade de analisar quais tabelas não são mais encontradas das que não continham quase-identificadores serve para avaliar o impacto que a anonimização teve na utilidade dos dados para busca por similares.

Percebemos, também, uma relação entre explicações e a análise dos resultados de uma busca por similares para fins de anonimização. Não basta apenas reconhecer quais tabelas são similares para se fazer uma análise sobre os dados, outras informações como "por que as tabelas são consideradas similares?", "quais colunas são consideradas similares?", "por que estas colunas são consideradas similares?" são necessárias para embasar a análise feita. Esta explicação mais detalhada permitiria também separar, do restante das tabelas, aquelas que são similares por causa de quase-identificadores. Mesmo na análise de uma segunda busca por similares, seria possível verificar se tabelas que estão em ambas foram encontradas com quase-identificadores ou não, e se há ou não risco dos quase-identificadores encontrados serem usados para ataques de ligação.

Como explicitado neste capítulo, não há consenso sobre explicações, tanto em relação à terminologia quanto em relação a formas de medir sua qualidade. Falta consenso, em particular, em calcular a qualidade de uma explicação, pois ela é muito dependente da definição do grupo

alvo e da motivação dessa explicação. No capítulo 3 apresentaremos nossa proposta de como estruturar uma explicação de forma que estas considerações sejam atendidas de modo mais satisfatório.

3 MODELO PROPOSTO

O capítulo 2 evidencia os diversos problemas relacionados à geração de explicações, especialmente no quesito de avaliar a qualidade dessas explicações. Desse modo, neste capítulo, propomos um modelo para construir explicações de forma que seja possível maximizar o benefício obtido.

Nesta seção mostramos o modelo proposto, com alguns exemplos de seu uso no caso mais específico da busca por similares, em seguida nas seções 3.1, 3.2, e 3.3 discutimos, de modo detalhado, os benefícios que nosso modelo pode trazer para as áreas de explicação, anonimização e busca por similares, respectivamente, na seção 3.4 apresentamos, em nível aprofundado, o fluxo que seguimos para montar a estrutura e utilizá-la para auxiliar a anonimização, por fim, na seção 3.5 discutimos algumas particularidades da implementação desenvolvida e apresentamos nossa implementação de um script em python3, que integra a construção da estrutura proposta em um script de busca por similares feito por outras pessoas.

Um dos principais motivadores na construção do modelo foi criar uma estrutura que pode ser adaptada facilmente para diversas situações em que explicações são necessárias, e que seja, também, facilmente adaptada para atender diversos grupos alvos de acordo com a necessidade de prover explicações.

A intenção é que a estrutura permita controlar como a explicação é apresentada, e que este controle sobre a apresentação facilite a geração de boas explicações, como isso pode ser feito é apresentado em mais detalhes na seção 3.1. Queremos usar a estrutura para separar a geração da explicação, da apresentação da explicação para humanos, dessa forma é possível focar em melhorar a apresentação sem se preocupar em como a busca por similares esta sendo feita, e vice-versa, já que a estrutura existiria como intermediário entre as duas frentes de trabalho.

A ideia é usar estruturas hierárquicas de dados para montar a explicação ao longo do processo. A título de exemplo, usamos uma representação em JSONs, mas a estrutura utilizada pode ser facilmente substituída por outras representações hierárquicas de dados. Escolhemos usar estas estruturas porque são modulares e, desse modo, podemos facilmente adicionar, remover ou modificar campos com partes da explicação. Esta modularidade permite, também, controlar de forma mais fácil a cobertura da explicação, já que é possível colocar cada parte do processo em um módulo diferente da estrutura.

Outro motivador para o uso de estruturas hierárquicas é que seu uso é comum em páginas web, e diversos analisadores (*parsers*) já existem para lidar com elas. Portanto, montar a explicação com essas estruturas permite que seja facilmente construída uma interface simples que apresente o resultado do processo e a explicação de forma mais legível, mesmo que a explicação tenha cobertura mais alta. Com uma interface um pouco mais complexa, é possível representar níveis diferentes de detalhamento na explicação para indivíduos com diferentes níveis de conhecimento sobre o processo.

Desse modo, a nossa proposta é que a construção dessa estrutura seja integrada no processo em que a explicação será necessária, cada componente do processo montando a explicação da sua própria função e agregando os resultados e as explicações das funções que chamou de acordo com a cobertura desejada.

A ordenação especifica da estrutura a ser usada é algo que deveria ser discutido dentro de cada área diferente em que as explicações se fariam necessárias. A figura 3.1 mostra uma versão simplificada de como a estrutura deve ser organizada para uma busca por similares.

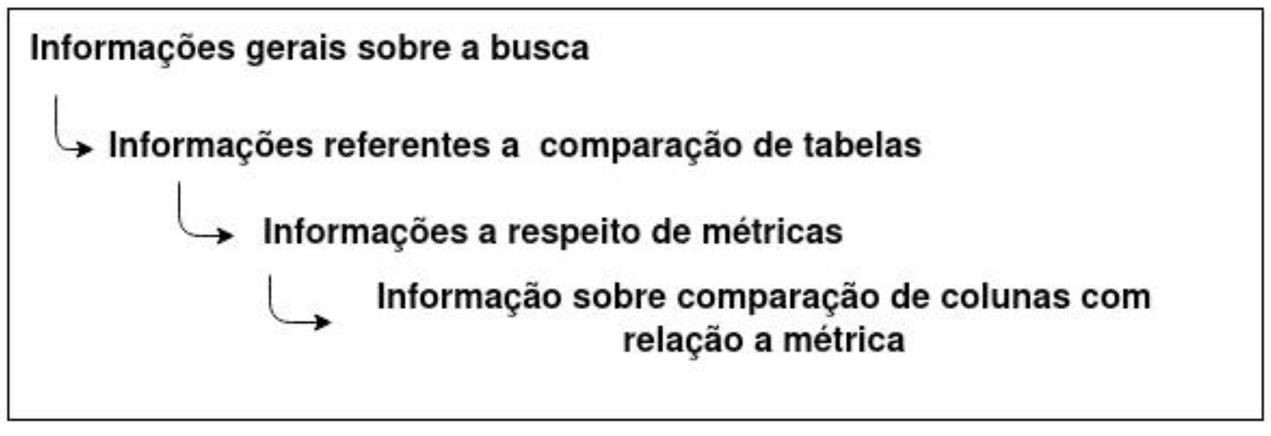

Figura 3.1: Estrutura simplificada

Apresentamos na listagem 3.1 um exemplo mais concreto de como a estrutura ficaria após ser o término do processo da busca por similares que a construiu.

```
1  {
2    "Tabela base": "Nome da tabela base"
3    "tabelas utilizadas": ["A","B",...,"Z"]
4    "Escolha de tabelas": "Como são escolhidas tabelas para comparar"
5    "Descrição da escolha": "Explicação mais detalhada sobre escolha"
6    "Comparações entre tabelas": [
7      {
8        "Tabelas comparadas": "tabela base" e "A"
9        "Conjunto de colunas da tabela base": [b_1, b_2, ... , b_X ]
10       "Conjunto de colunas de A": [ A_1, A_2, ... , A_Y ]
11       "Similaridade entre tabelas": [
12         {
13           "Métrica" : "Nome da métrica",
14           "Como funciona" : "Explicação sobre como a métrica funciona",
15           "Valor" : "Similaridade entre as tabelas para a métrica",
16           "Comparações entre colunas": [
17             {
18               "Colunas comparadas": [b_w, A_z],
19               "N linhas coluna base": |[b_w]|
20               "N linhas coluna comparada": |[A_z]|
21               "Valor": "Similaridade entre as colunas para a métrica"
22             }, ...
23           ]
24         },
25         ...
26       ]
27     }, ...
28   ]
29 }
```

Listagem 3.1: Exepmlo da estrutura

Usando o esquema do modelo presente na listagem 3.1 como referência, a parte da estrutura referente às linhas 2 a 5 seria preenchida como parte da inicialização da busca, preenchendo as informações mais gerais sobre a busca que estiver sendo realizada. Já na linha 6, temos um vetor que contém as informações de comparações entre tabelas, a comparação entre tabelas propriamente dita seria executada em uma função separada, retornando a subestrutura que contém a comparação entre duas tabelas para cada par de tabelas comparadas.

O conteúdo das linhas 7 a 11 refere-se à subestrutura que contém informações sobre a comparação de duas tabelas; as linhas 8 a 10 apresentam informações mais gerais sobre a comparação como os nomes das tabelas e as colunas comparadas; já a linha 11 contém um vetor com a subestrutura que contém a comparação propriamente dita, cada elemento do vetor representa uma forma diferente de comparar o par de tabelas.

Nas linhas 12 a 16 estão as informações sobre a comparação entre tabelas para uma determinada métrica. As linhas 13 a 15 informam a respeito da métrica utilizada e o valor de similaridade encontrado utilizando essa métrica; a linha 16 contém um vetor com as subestruturas que representam as comparações entre colunas que foram utilizadas para calcular a similaridade entre tabelas.

Por fim, as linhas 17 a 22 mostram um exemplo da subestrutura de comparação entre colunas, contendo os nomes das colunas, a quantidade de elementos de cada coluna e o valor encontrado na comparação entre elas.

Usando estruturas com subestruturas aninhadas é mais fácil modificar diferentes partes da estrutura, assim como adicionar e remover campos conforme necessário. Por exemplo, se for necessário ter um campo com o tempo que a comparação levou (tanto para a comparação individual entre um par de colunas quanto para a comparação entre tabelas) é extremamente fácil adicionar o campo na estrutura. Da mesma forma, é fácil remover campos caso seja determinado que são desnecessários.

A intenção é que a estrutura de explicação gerada seja armazenada em um arquivo diferente do resultado da busca por similares (ou qualquer outra aplicação que utilize nossa estrutura) de modo que seja possível obter os resultados da busca por similares de forma rápida quando necessário, e avaliar a explicação gerada de forma separada. Não é nossa intenção limitar o uso de ferramentas existentes em função da estrutura de explicação que propomos

3.1 BENEFÍCIOS PARA EXPLICAÇÃO

Um dos benefícios de usar JSONs para montar explicações é que eles são frequentemente usados em páginas web, e, por isso, existem diversas ferramentas para lidar com este tipo de estrutura. Além disso, com uma página web simples é possível se utilizar de fatores que melhoram a qualidade de uma explicação sem ter que modificar a explicação em si. Mesmo que não seja necessário o uso de uma página web para lidar com a estrutura, o simples fato de que a estrutura é padronizada facilita criar *scripts* para tratar a estrutura e fazer as análises necessárias.

O modelo de uso que propomos assume que a explicação pode ter cobertura alta, mesmo que a custo de legibilidade do resultado. A partir da estrutura proposta, é possível separar facilmente segmentos diferentes da explicação na hora de apresentá-los em uma página web, de forma que, para cada grupo alvo diferente conhecido, é possível apresentar a mesma explicação de formas diversas.

Além disso, colocar a explicação em uma página web simples permite que possamos instigar a curiosidade de quem recebe a explicação. Para fazer isto, basta que ao invés de apresentar a explicação inteira de uma vez, apresentemos apenas parte do resultado e adicionemos algo para forçar um usuário a interagir com a página, como um botão, por exemplo, para receber mais partes da explicação. A interação forçada coloca mais agência no receptor da explicação, ele passa de apenas receptor de explicação para aquele que se mobiliza para encontrar respostas.

Outro benefício é evitar sobrecarga de informações desnecessárias. Esse benefício é decorrente do controle da quantidade de elementos de uma explicação, pois ele evita que um usuário tenha que lidar com informações que não sejam relevantes em determinados contextos. Não é nem mesmo necessário que esse particionamento aconteça de forma linear, já que as dúvidas que um usuário pode formular a partir de um pedaço de explicação variam de acordo com sua curiosidade.

Usando como exemplo a estrutura na listagem 3.2 temos as informações relevantes do resultado de uma busca por similares, com a possível exceção dos conteúdos das tabelas, que podem ser obtidos caso seja necessário. A partir dessa estrutura, podemos montar diversas

formas de apresentar o resultado encontrado, essa apresentação fica submetida ao que é mais relevante para quem fez a busca por similares.

Podemos, por exemplo, restringir a mostra a apenas o nome da tabela encontrada, aos valores de similaridade calculados. Detalhes, como a forma de calcular a similaridade, as métricas utilizadas para chegar ao valor de similaridade entre colunas e os valores de similaridades entre colunas, podem ser escondidos de quem efetuou a busca por similares até que haja alguma interação com a interface que peça estas informações, uma de cada vez.

```
{
  "Tabela base": "Tx"
  "tabelas utilizadas": ["A"]
  "Escolha de tabelas": "Escolha unica"
  "Descrição da escolha": "compara apenas com a primeira tabela encontrada"
  "Comparações entre tabelas": [{
     "Tabelas comparadas": "Tx" e "A"
     "Conjunto de colunas da tabela base": [x]
     "Conjunto de colunas de A": [ a ]
     "Similaridade entre tabelas": [ {
        "Métrica" : "Similaridade de cosseno",
        "Como funciona" : "https://en.wikipedia.org/wiki/Cosine_similarity",
        "Valor" : "0.9"
        "Comparações entre colunas": [{
           "Colunas comparadas": [x,a],
           "N linhas coluna base" : 15000
           "N linhas coluna comparada" : 20000
           "Valor": "0.9"
        }]
     } ]
  } ]
}
```

Listagem 3.2: Exemplo de uso do modelo

Outra possibilidade é que sejam utilizadas diversas métricas, tanto para calcular as similaridades entre colunas, quanto para combinar estas similaridades em uma similaridade entre tabelas. Neste caso, outra rota que pode ser tomada é apresentar, antes dos resultados da busca, uma escolha de qual técnica o usuário deseja ver, fazendo com que o próprio começo da interação seja diferente, isso pode interferir nas possíveis dúvidas iniciais que um usuário teria. Desse modo, a estrutura proposta permite que se desconsidere a forma como uma explicação será apresentada e concentre-se em uma maneira que seja adequada para o contexto em que a explicação é utilizada.

Precisamos reconhecer, no entanto, que o uso de uma estrutura hierárquica de dados para construir e apresentar uma explicação não resolve problemas como a falta de consenso na terminologia utilizada para explicações, nem resolve o problema de não existirem métricas automáticas para medir a qualidade de uma explicação (e como apontado por Hoffman et al. (2018) não é desejável que a qualidade de uma explicação seja medida por uma métrica descontextualizada). Porém, utilizar uma estrutura hierárquica de dados permite prover uma explicação de forma que os fatores considerados como parte de uma explicação boa sejam impulsionados, enquanto os fatores que podem prejudicar uma explicação são mitigados.

3.2 BENEFÍCIOS PARA ANONIMIZAÇÃO

Como discutido na seção 2.4 é possível utilizar a busca por similares para fazer uma análise de risco mais informada antes de anonimizar os dados, e também fazer uma análise da

qualidade da anonimização realizada comparando o resultado da busca inicial com o resultado de uma busca feita com a mesma tabela após os dados terem sido anonimizados.

Utilizar a estrutura proposta facilita esta análise, deixando mais evidente quais foram as colunas comparadas e quais foram as colunas mais relevantes para o resultado da busca por similares. Essa utilização permite que, na comparação entre as buscas realizadas pré e pós anonimização, seja possível afirmar se um quase-identificador influenciou o resultado, ou se uma tabela foi encontrada em ambas as buscas porém na segunda busca não possuía atributos similares considerados quase-identificadores.

Para exemplificar o processo, utilizaremos a tabela 3.1 como a tabela a ser anonimizada (EX), e a tabela 3.2 como uma das tabelas encontradas durante a busca (A). Na listagem 3.3 , expomos como a estrutura apresenta o resultado da busca por similares; os valores apresentados são meramente ilustrativos, servem apenas para exemplificar como vemos a estrutura sendo utilizada.

CPF	UF	Condição
123	PR	COVID
456	SP	Pancreatite
789	TO	Ulceras

Tabela 3.1: Tabela base, "EX"

CPF	UF	Nome
123	PR	Fulano
456	SP	Ciclano
789	RS	Beltrano

Tabela 3.2: Tabela encontrada, "A"

```
{
  "Tabela base": "EX"
  "tabelas utilizadas": ["A"]
  "Escolha de tabelas": "Comparação com todas"
  "Descrição da escolha": "Compara todas as tabelas encontradas"
  "Comparações entre tabelas": [{
     "Tabelas comparadas": "EX" e "A"
     "Conjunto de colunas da tabela base": [CPF, UF, Condição]
     "Conjunto de colunas de A": [ CPF, UF, Nome ]
     "Similaridade entre tabelas": [{
        "Métrica" : "Média da Similaridade de cosseno",
        "Como funciona" : "Média das similaridades de cosseno para as
            colunas comparadas;
            https://en.wikipedia.org/wiki/Cosine_similarity",
        "Valor" : "0.7"
        "Comparações entre colunas": [{
           "Colunas comparadas": [CPF,CPF],
           "valor": "1.0"
         },{
           "Colunas comparadas": [UF,UF],
           "valor": "0.8"
         }, {
           "Colunas comparadas": [Condição,Nome],
           "valor": "0.3"
       }]
     }]
  }]
}
```

Listagem 3.3: Exemplo do modelo em uso para anonimização parte 1

Uma análise da estrutura mostra que as tabelas tiveram uma alto grau de similaridade, além disso, fica evidente, também, que as colunas que mais influenciaram o resultado foram

CPF e UF, já que a coluna "Condição" não foi considerada muito similar com sua possível contrapartida na tabela "A". Uma entidade que queira publicar a tabela "EX", de forma anônima, consegue então decidir como anonimizar a tabela, considerando a importância da coluna "CPF" e como essa coluna poderia ser utilizada é decidido que ela precisa ser tratada. Por outro lado, como a coluna "UF" não revela informações que a entidade considera importante, ela não será alterada. A tabela anonimizada pode ser vista em 3.3 e a estrutura construída com uma segunda busca por similares pode ser vista em 3.4.

CPF	UF	Condição
1**	PR	COVID
4**	SP	Pancreatite
7**	TO	Ulceras

Tabela 3.3: Tabela anonimizada

Analisando a nova estrutura gerada é fácil verificar que as tabelas já não são mais consideradas similares, e que a coluna que tinha sido definida como quase-identificador não pode mais ser ligada tão facilmente com colunas que anteriormente seriam consideradas similares. Ao mesmo tempo, a informação carregada pela coluna "UF", que não tinha sido considerada como um risco para a anonimização, reteve sua relevância para comparações de "UF".

```
{
  "Tabela base": "EX"
  "tabelas utilizadas": ["A"]
  "Escolha de tabelas": "Comparação com todas"
"Descrição da escolha": "Compara todas as tabelas encontradas"
"Comparações entre tabelas": [{
   "Tabelas comparadas": "EX" e "A"
   "Conjunto de colunas da tabela base": [CPF, UF, Condição]
   "Conjunto de colunas de A": [ CPF, UF, Nome ]
   "Similaridade entre tabelas": [{
      "Métrica" : "Média das similaridades de cosseno",
      "Como funciona" : "Média das similaridades de cosseno para as
         colunas comparadas;
         https://en.wikipedia.org/wiki/Cosine_similarity",
      "Valor" : "0.37"
      "Comparações entre colunas": [{
         "Colunas comparadas": [CPF,CPF],
         "valor": "0.01"
       },{
         "Colunas comparadas": [UF,UF],
         "valor": "0.8"
       }, {
         "Colunas comparadas": [Condição,Nome],
         "valor": "0.3"
      }]
   }]
 }]
}
```

Listagem 3.4: Exemplo do modelo em uso para anonimização parte 2

Caso a entidade anonimizando a tabela ainda não esteja satisfeita com o nível de proteção provida, pode repetir o processo de busca por similares, analise de riscos e anonimização até que esteja satisfeita com o resultado obtido, apresentamos isso de forma mais detalhada na seção 3.4.

3.3 BENEFÍCIOS PARA BUSCA POR SIMILARES

No caso da busca por similares, os benefícios estão relacionados com a existência da explicação em si, a estrutura em si traz apenas os benefícios para explicações já discutidos em 3.1.

Explicar o processo de busca por similares permite que a analise dos resultados seja efetuada mais facilmente, e com mais qualidade, já que a explicação provém fundamentação para as similaridades encontradas.

A explicação permite também que o desenvolvimento e melhoramento de ferramentas de busca por similares ocorra mais rapidamente, pois é possível acompanhar o processo e identificar problemas mais facilmente.

Além disso, em cenários em que a busca por similares (ou integração), é utilizada para gerar indicadores que serão publicados, a explicação provém mais informações para serem usadas pelos indicadores. A explicação permite também manter a rastreabilidade das consultas efetuadas, auditar a geração dos indicadores, e validar o processo de busca por similares e a construção de indicadores.

3.4 PROCESSO

Nesta seção, apresentaremos um esquema com o fluxo de execução que seguimos para montar a estrutura de explicação utilizada e explicaremos como aproveitar-se dela para refinar o processo de anonimização. A figura 3.2 apresenta um esquema simplificado de como a busca explicada funciona, e a figura 3.3 mostra o fluxo que será discutido ao longo da sessão.

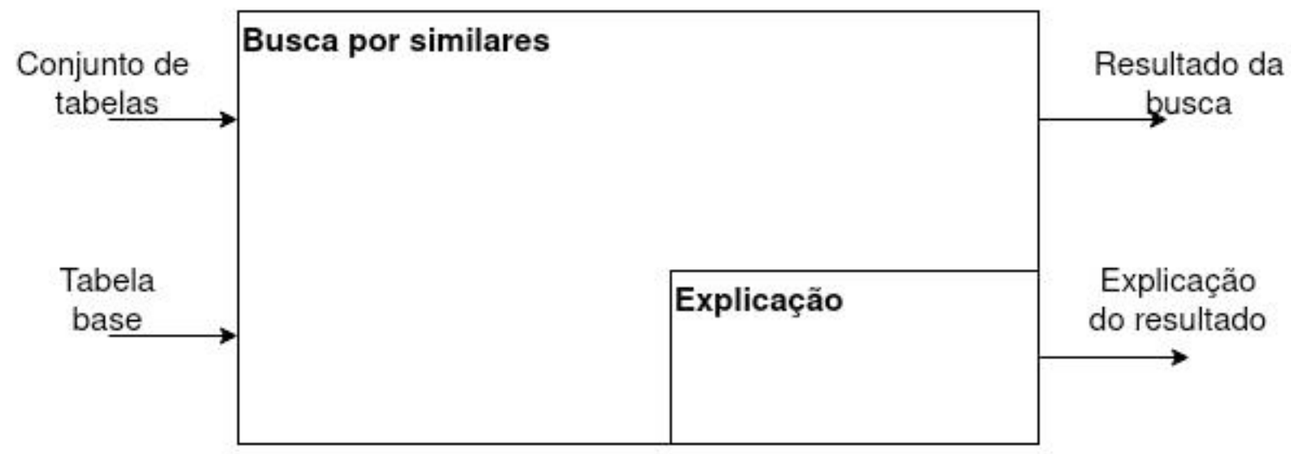

Figura 3.2: Esquema simplificado

Evidentemente, a primeira coisa a ser feita é definir qual é a tabela será usada como base de comparação de similaridade durante a busca, e obter um conjunto de tabelas em que a busca será realizada. Os trabalhos da área, apresentados na seção 2.1, utilizaram tabelas armazenadas em grandes repositórios abertos, como *EU Open Data*, e *Project Open Data*, porém, nós achamos a api provida difícil de utilizar. Logo, por questão de praticidade, utilizamos o conjunto de tabelas do banco **Dados Educacionais** presente na C3SLdatabase.

Em seguida, foi necessário selecionar uma tabela do conjunto para ser comparada com a tabela base. Esta seleção pode ser feita de diversas maneiras, algumas delas foram discutidas em alguns dos trabalhos apresentados em 2.1, para otimizar a seleção e reduzir o número de comparações que são necessárias, no nosso caso, como estamos trabalhando com um conjunto relativamente pequeno de tabelas (87 contando com a tabela base) decidimos iterar por todas as tabelas. A cada passo do processo, as informações relevantes são adicionadas na estrutura de explicação da busca.

Uma vez que a tabela a ser comparada tenha sido selecionada, é necessário calcular a similaridade entre as colunas das duas tabelas, e escolher que pares de colunas serão considerados

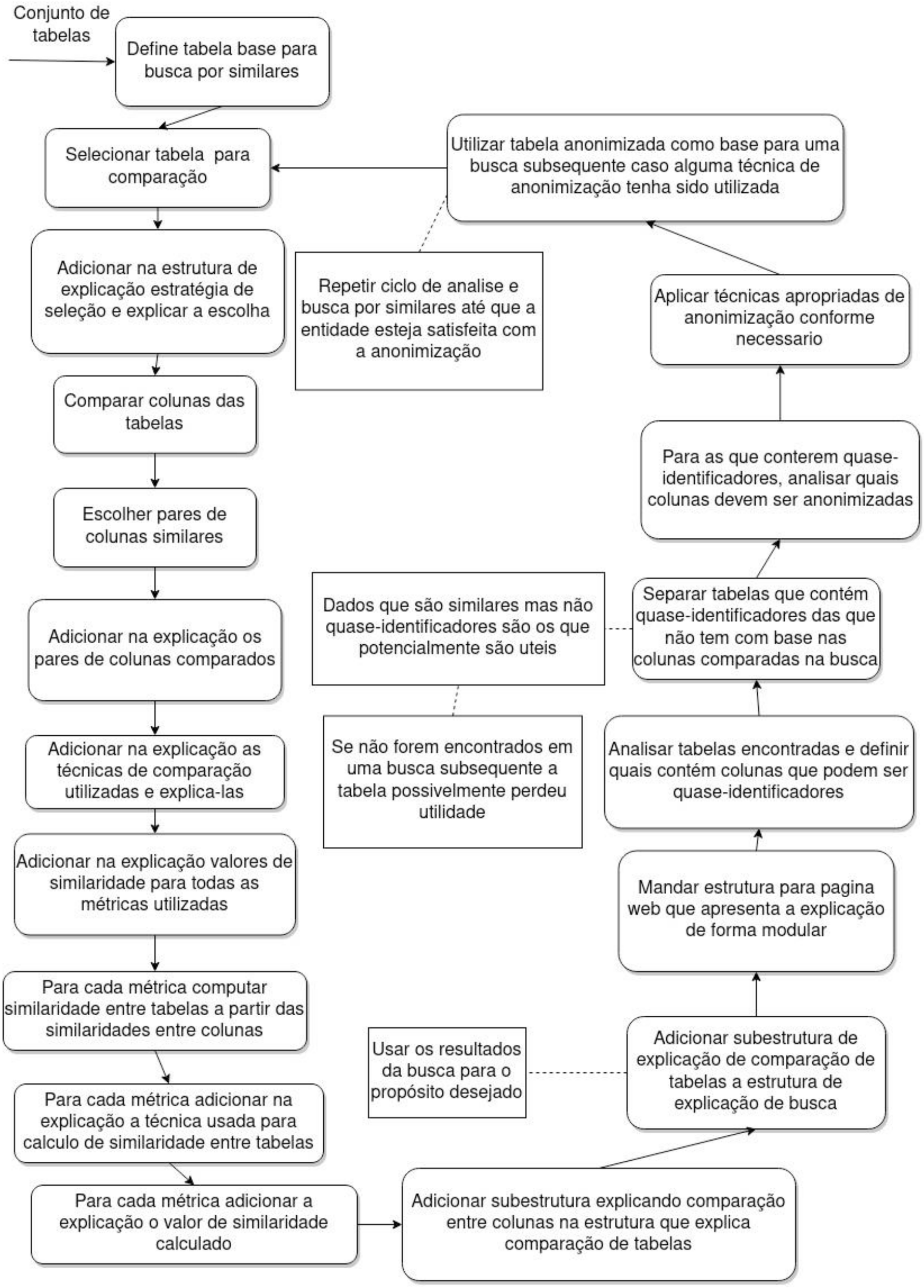

Figura 3.3: Fluxo

similares; para cada coluna da tabela base pode haver apenas uma coluna considerada similar na tabela comparada. Para obter este conjunto de pares de colunas similares, escolhemos comparar todos os pares de colunas passíveis de obtenção; utilizamos essas colunas como vértices e as similaridades calculadas como peso para as arestas, a fim de montar um grafo completo. Em

seguida, calculamos o pareamento maximal deste grafo para obter os pares de colunas que maximizam a similaridade entre as colunas.

Um grafo diferente é gerado para cada métrica de comparação utilizada durante a busca e o resultado de cada pareamento é armazenado separadamente. Além disso, para cada métrica utilizada, um conjunto diferente de pares de colunas é armazenado na estrutura após o pareamento. Uma das desvantagens de usar um grafo para fazer estes cálculos é que ele limita a quantidade de pares de colunas de acordo com a tabela que tiver menos colunas.

Em seguida, é necessário combinar as comparações entre colunas em um valor de similaridade entre tabelas. Para fazer essa combinação, decidimos utilizar uma média simples entre os pares de colunas considerados similares. Essa média é feita de forma separada para cada métrica de similaridade utilizada. Outras formas de fazer este cálculo podem ser facilmente adicionadas à estrutura conforme necessário.

Os valores calculados para cada métrica de similaridade, tanto entre colunas (após o pareamento) quanto entre tabelas, são armazenados no nível apropriado da estrutura (essa explicação está detalhada na seção 3.5). A estrutura final de comparação entre as duas tabelas é salva em um vetor de comparações, isso se repete até que todas as comparações tenham sido feitas.

Como a estrutura e o resultado da busca podem ser armazenados em arquivos separados, após o término desse processo, os resultados podem ser utilizados independentemente da estrutura de explicação, caso ela não seja necessária. Por outro lado, como no caso de auxiliar a anonimização de dados, a estrutura pode ser utilizada para melhorar a análise dos resultados.

Como mencionado em 3.1, é possível utilizar a estrutura de explicação construída em uma página web de forma a controlar melhor o *display* das informações explicando os resultados da busca por similares.

Uma vez que o analista tenha em mãos a estrutura, ele pode averiguar quais das tabelas similares são realmente similares, quais contém quase-identificadores e quais são falsos positivos. Com isso, o analista consegue agrupar as tabelas separadamente: reunindo-as em um conjunto das que trazem risco para a anonimização e em um outro em que não oferecem esse risco.

A partir do conjunto de tabelas que trazem risco, é possível tomar uma decisão mais informada a respeito de como anonimizar a tabela base de forma a mitigar o risco de quebra de anonimização.

Com a tabela anonimizada, é possível utilizá-la como base para realizar uma segunda busca por similares. O processo seguido é praticamente o mesmo, com a diferença de que, quando forem obtidos e identificados os conjuntos de tabelas que trazem risco e os que não trazem, é possível fazer uma segunda análise dos resultados. Caso uma tabela que estava no grupo de risco persistir durante a segunda busca por similares, tem-se um indicador de que a anonimização efetuada pode não ter sido suficiente. Tabelas que não foram encontradas durante esta segunda busca podem indicar perda de utilidade dos dados, já que menos informações são conectadas à tabela anonimizada.

Idealmente, este ciclo de anonimização e análise seria repetido até que a entidade que estivesse publicando os dados ficasse satisfeita com a anonimização provida. O processo pode, inclusive, ser efetuado em diversos repositórios de dados de forma a dar ainda mais embasamento para os resultados da análise.

3.5 IMPLEMENTAÇÃO

Nesta seção, apresentaremos o script python implementado para a realização dos experimentos que serão expostos na seção 4.1. De maneira geral, o script conecta-se em uma

base de dados, obtém informações sobre as tabelas no banco (nomes das tabelas, das colunas de cada tabela e os dados contidos nas colunas), aponta uma das tabelas - escolhida de forma arbitrária por nós - para ser a tabela de base da busca, e itera por todas as tabelas comparando-as e montando a estrutura de explicação. Ao longo dessa seção, aprofundaremos como cada parte do script funciona.

A conexão com o banco de dados é feita com pymonetdb, uma API para interagir com bases de dados monetdb. Após estabelecer essa conexão, fazemos uma consulta simples para obter o nome de todas as tabelas do banco que não pertencem ao sistema do banco ou que são views. Assim, podemos construir a parte mais externa da estrutura de explicação, que apresenta o nome da tabela base, expõe as tabelas usadas na busca e mostra como foi a escolha de qual das tabelas comparar. Esta parte da estrutura pode ser vista na listagem 3.5.

```
1  my_struct = {
2      "Tabela base":' '.join(tabelas[X]),
3      "Tabelas comparadas na busca":[tabelas],
4      "Escolha de tabelas":"Técnica de comparação X",
5      "Comparações entre tabelas": []
6  }
```

Listagem 3.5: Exemplo de parte mais externa da estrutura de explicação

Em seguida, iteramos por todas as tabelas obtidas (87), e comparamos com a tabela base (ela também é comparada para termos um ponto de referência para as técnicas de comparação utilizadas nesse processo). Essa comparação é feita em uma função que tem como retorno a subestrutura de comparação de tabelas com todas as suas informações preenchidas. A subestrutura é então adicionada ao vetor de comparações entre tabelas realizadas e a próxima tabela é comparada. Finalizado esse processo, o resultado final é salvo em um arquivo JSON, que pode ser utilizado para fazer as análises desejadas.

Inicialmente, a comparação entre tabelas fornece as colunas que cada tabela possui e, a partir dessas informações, o script preenche, tantos campos da estrutura de comparação de tabelas quanto possível. Essa estrutura inicial pode ser observada na listagem 3.6. Em seguida, usamos as informações disponíveis e fazemos as comparações entre as colunas. Essas comparações são realizadas em uma função separada que retorna um vetor, cada posição do vetor contém uma subestrutura de comparações de colunas, cada subestrutura contém as comparações feitas com uma métrica diferente. A correspondência entre posição e métrica foi uma decisão arbitrária.

```
1  my_struct = {
2      "Tabelas Comparadas":tabela_base+', '+tabela_comparada,
3      "Colunas Tabela base":colunas_t1,
4      "Colunas Tabela comparada":colunas_t2,
5      "Similaridade entre tabelas":[],
6  }
```

Listagem 3.6: Subestrutura referente a uma comparação entre tabelas

Este vetor é então passado para uma outra função que combina as similaridades entre colunas em similaridade entre tabelas. Escolhemos fazer uma média simples das similaridades calculadas para cada técnica de comparação utilizada com a adição de uma métrica adicional que é a média das 4 similaridades encontradas. A função que faz isso pode ser encontrada na listagem 3.7. O modo como a subestrutura de similaridade entre tabelas é construída permite que mais formas de comparação sejam facilmente adicionadas, basta que sejam adicionadas no vetor e que a função de cálculo de similaridade seja adaptada para lidar com as posições novas do vetor.

```
def table_sim(comps):
    avg_cos = 0
    avg_jacc = 0
    avg_cscore = 0
    avg_cdf = 0
    size = len(comps[0])
    for c in comps[0]:
        if(c["Valor"] >= 0):
            avg_cos = avg_cos + c["Valor"]
    if(size > 0):
        avg_cos = avg_cos / size
    size = len(comps[1])
    for c in comps[1]:
        if(c["Valor"] >= 0):
            avg_jacc = avg_jacc + c["Valor"]
    if(size > 0):
        avg_jacc = avg_jacc / size
    size = len(comps[2])
    for c in comps[2]:
        if(c["Valor"] >= 0):
            avg_cscore = avg_cscore + c["Valor"]
    if(size > 0):
        avg_cscore = avg_cscore / size
    size = len(comps[3])
    for c in comps[3]:
        if(c["Valor"] >= 0):
            avg_cdf = avg_cdf + c["Valor"]
    if(size > 0):
        avg_cdf = avg_cdf / size
    avg_all = (avg_cos + avg_jacc + avg_cscore + avg_cdf)/4
    return([
    {"Metrica":"Distancia de cosseno", "explicacao":"",
        "Valor":avg_cos,"Comparação colunas":comps[0]},
    {"Metrica":"Distancia de jaccard", "explicacao":"",
        "Valor":avg_jacc,"Comparação colunas":comps[1] },
    {"Metrica":"Containment score", "explicacao":"",
        "Valor":avg_cscore,"Comparação colunas":comps[2] },
    {"Metrica":"Distribuiçao cumulativa", "explicacao":"",
        "Valor":avg_cdf,"Comparação colunas":comps[3] },
    {"Metrica":"Media geral", "explicacao":"","Valor":avg_all,"Comparação
        colunas": [] }] )
```

Listagem 3.7: Função de cálculo de similaridade entre tabelas a partir da similaridade entre colunas

A função que calcula a similaridade entre colunas é simples, ela itera por todos os pares possíveis de colunas que podem ser comparados. A cada iteração, armazena os dados contidos nas colunas em vetores, e para cada forma de comparação utilizada monta a subestrutura de comparação de colunas, salvando a subestrutura para cada técnica de comparação em um vetor separado. Um exemplo da subestrutura pode ser observado na listagem 3.8 .

```
col_sim = {
    "Colunas comparadas" : [ColA, ColB],
    "N linhas base" : |ColA|,
    "N linhas comparada" |ColB|,
    "Valor" : sim_entre_colunas
}
```

Listagem 3.8: Subestrutura de comparação entre um par de colunas

Uma vez que todas as colunas tenham sido comparadas, cada vetor de comparações (reiterando que cada forma de comparação é armazenada em um vetor de similaridade de colunas separado) é enviado para uma função que reduz o número de similaridades a serem consideradas. Isso pode ser feito a partir da criação de um grafo bipartido, em que os nodos são as colunas e as arestas com pesos são as similaridades calculadas, fazemos então um pareamento maximal deste grafo para obter os pares de colunas que são considerados mais similares. Como cada forma de comparação é tratada separadamente, os grafos finais não necessariamente serão os mesmos, por isso, é importante que a estrutura de explicação informe quais são os pares finais de colunas ao final do processo.

Em termos de implementação, para fazer um pareamento maximal do grafo, fazemos um pareamento mínimo, invertendo (1 - similaridade) o valor de similaridade calculada. Para melhorar os resultados, removemos pares de colunas com similaridade abaixo de um limiar arbitrário, porém, para evitar erros quando a remoção não deixasse arestas o suficiente para realizar o pareamento, ao invés de remover as arestas, elas receberam o peso máximo para similaridades (como todas as métricas utilizadas usam valores no intervalo [0,1], o peso máximo é 1).

Entre os modelos de comparação, escolhemos utilizar: similaridade de cosseno, similaridade de jaccard, score de containment e o cálculo cumulativo das probabilidades de ambas as colunas pertencerem ao mesmo domínio. A escolha desses modelos está fundamentada nos estudo que embasam essa pesquisa. Os limiares utilizados para cada métrica foram respectivamente 0.3, 0.4, 0.4 e 0.4 definidos empiricamente com testes de diferentes valores.

Além disso, para diferenciar casos em que a similaridade calculada foi 0 de casos em que era impossível calculá-la, usamos uma marcação de erro. Quando não é possível fazer o cálculo de similaridade, o valor é marcado como -1, para todas as contas realizadas, isso foi tratado como um 0, a marcação -1 é relevante apenas na estrutura de explicação.

As situações em que a comparação não é possível (ou não traz informação nenhuma) incluem casos em que a coluna contém dados booleanos, casos em que a coluna está vazia e casos em que o esvaziamento da coluna ocorre durante o pré-processamento dos dados. O pré-processamento é aplicado às colunas antes da comparação (exceto no caso em que ambas as colunas contêm números inteiros) para normalizar os dados, todos os caracteres são colocados em minúsculo, removendo caracteres desconhecidos e especiais, além de palavras que não estão no vocabulário conhecido pelo modelo pré-treinado de *Word 2 Vec*. Utilizamos o modelo em português provido por *fasttext*.

3.6 CONSIDERAÇÕES

Após diversas leituras, identificamos múltiplas lacunas nas áreas de busca por similares, anonimização e explicação. Neste capítulo, apresentamos nossa proposta para resolver, ou, no mínimo, amenizar, os problemas trazidos por essas lacunas.

Na área de anonimização, há o risco de ataques de inferência levando à identificação de indivíduos. Esse risco pode ser amenizado com a realização de uma busca por similares em repositórios de dados abertos com a tabela a ser anonimizada, isso permitirá que seja feita uma análise de riscos mais completa.

Reconhecemos que apenas a busca por similares não é ideal, já que deixa todo o processo de análise por conta da entidade que está anonimizando os dados. Para melhorar este processo, é necessário adicionar uma explicação ao resultado da busca, a fim de facilitar a verificação e dar mais embasamento para análise de riscos feita pela entidade que estaria anonimizando os dados.

A definição de uma boa explicação é uma tarefa bastante complexa. Há empecilhos de diversos fatores, muitos deles relativos ao grupo alvo que vai recebê-la. Nossa solução para tentar fornecer uma boa explicação, apesar de toda a complexidade, foi produzir uma estrutura padronizada e modular que pode ser usada não só para construir explicações com diferentes níveis de detalhamento na estrutura, mas também que possa ser facilmente analisada e tratada com *parsers* comuns.

A ideia é que, com uma estrutura padrão tão fácil de construir como de manipular, podemos escolher, sem complicações, como e quanto da explicação apresentar. A intenção é que seja possível maximizar os fatores que melhoram a qualidade de uma explicação, ao mesmo tempo em que minimizamos aqueles que a pioram, a partir de manipulações simples na hora de apresentar as informações contidas na estrutura de explicação.

Com uma explicação boa do processo de busca, é possível então fazer uma análise mais rigorosa dos riscos de desanonimização de uma base de dados e tomar decisões claras em relação a quais técnicas empregar para manter a privacidade dos dados.

4 EXPERIMENTOS

Neste capítulo, apresentaremos os experimentos realizados a partir dos estudos desenvolvidos para este trabalho. Ele se divide em três seções, na seção 4.1, descrevemos os experimentos realizados. Na seção 4.2, apresentamos os resultados obtidos. Por fim, na seção 4.3, discutimos as conclusões a que chegamos a partir dos resultados encontrados.

4.1 EXPERIMENTOS

Para a realização dos experimentos, foi utilizado como base o script python, fornecido por Maria Helena Franciscatto, aluna de doutorado do professor Marcos Didonet del Fabro, do departamento de informática da UFPR. Esse script realiza uma busca por similares parcial, parando no cálculo da similaridade entre colunas. A partir desse ponto, inserimos algumas modificações como completar o processo de busca por similares, combinando as similaridades entre colunas em similaridades entre tabelas, integramos a geração da estrutura de explicação no processo de busca por similares, e corrigimos alguns *bugs* encontrados durante a implementação.

De modo resumido, o experimento consistiu em uma execução do processo descrito na seção 3.4. A base de dados utilizada para o experimento foi banco "Dados Educacionais" da *c3sldatabase*, que contém 87 tabelas (excluindo-se tabelas que existem apenas para propósito de mapeamento e tabelas de sistema interno do banco), armazenadas em uma instância do MonetDB. Esta base de dados foi escolhida pela facilidade de acesso, além do fato de que, pelos dados tratarem de assuntos similares, já esperávamos um certo grau de similaridade entre eles.

Os dados contidos na base de dados "Dados educacionais" dizem respeito a diversos programas de auxílio social do governo, como o Fies(Fundo de Financiamento ao Estudante do Ensino Superior), e diversos censos sócio-econômicos, como a PNAD(Pesquisa Nacional por Amostra de Domicílios) e SIM(Sistema de Informações sobre Mortalidade). Essas informações podem ser obtidos de forma aberta nos diversos sites do governo. A base "Dados educacionais" apenas centraliza e padroniza estes dados de modo que sejam mais fáceis de serem utilizados e acessados.

Devido a limitações no ambiente de teste, não foi possível realizar os experimentos com os dados completos. Para contornar este problema, decidimos então separar o experimento em dez amostras, cada uma delas contendo no máximo 1500 elementos, selecionados com a função "*SAMPLE*" do MonetDB, que faz uma seleção de elementos com uma distribuição uniforme e sem reposição.

Para executar o processo da seção 3.4 escolhemos, de forma arbitrária, uma das tabelas de um conjunto de 87 — que foi mantida em todas as amostras – para servir de base para a busca por similares. Como estávamos lidando com um conjunto pequeno de tabelas, não achamos necessário empregar nenhuma técnica avançada de seleção de tabelas a serem comparadas, optando por simplesmente iterar sobre o conjunto delas. A tabela base foi comparada com si mesma para servir de controle, tanto durante o teste da implementação da busca por similares, quanto no teste após tê-la anonimizado.

Após o término da primeira leva de buscas por similares, fizemos uma análise dos resultados, avaliando as similaridades obtidas e utilizando a estrutura para auxiliar na análise dos resultados. Com base nessa análise, escolhemos quatro atributos da tabela base para serem anonimizados. A técnica utilizada para anonimizar estes atributos foi a supressão, uma das mais

simples possível. Essa técnica consiste em deixar o atributo na tabela, porém sem nenhum valor válido em nenhum dos elementos.

Assim que a anonimização da tabela foi concluída, fizemos novamente a busca por similares nas dez amostras e analisamos os resultados para ver como as similaridades calculadas foram afetadas pela anonimização.

Todos os experimentos foram executados em uma das máquinas virtuais providas pelo DINF (Departamento de Informática). A máquina virtual utilizada possui 15 GB de memória e 16 cpus do modelo "Common KVM processor".

4.2 RESULTADOS

Após a primeira leva de execuções, percebemos que, das métricas utilizadas (similaridade de cosseno, similaridade de jaccard, *containment score* e distribuição cumulativa) apenas a similaridade de cosseno obteve valores acima de 0.5. Esse resultado era esperado, já que os dados da tabela de base eram, em sua maioria, texto e, das quatro métricas, a similaridade de cosseno é a que melhor funciona para texto, enquanto as outras métricas são mais adequadas para dados numéricos.

Das 87 tabelas, apenas 26 tiveram sua similaridade acima de 0.5, dessas, apenas quinze ficaram acima de 0.6. Desse último conjunto, apenas dois obtiveram o valor máximo de similaridade de 1.0, a saber, a tabela de base propriamente dita (nosso grupo controle), e outra tabela que agregava dados a mais na tabela base (como a estratégia de comparação favorece tabelas com menos colunas na comparação, as colunas extras da agregação acabaram sendo ignoradas).

Os valores de similaridades encontrados podem ser observados na tabela 4.1, o sinal de asterisco, "*", marca a tabela que serviu como base . Esses valores representam uma média das dez amostras e o desvio padrão das similaridades entre as diferentes amostras. "Cos" é a similaridade de cosseno, "Jacc" é a similaridade de Jaccard, "Cscore" é a medida de *containment*, "CDF" é a métrica de probabilidade cumulativa e "Média" é uma média das médias de similaridade para as funções utilizadas (em cada amostra calculamos a média entre as similaridades calculadas, e aqui apresentamos a média destas médias), as colunas que começam com "σ" são as colunas que contêm o desvio padrão de similaridade para a devida métrica.

Como é possível observar na tabela 4.1, com exceção da métrica de similaridade de cosseno, as outras métricas apresentaram performances pobres para todas as tabelas, com exceção da tabela base e da tabela agregada a partir da tabela base.

Isto se deve, principalmente, a dois fatores: o primeiro refere-se ao fato de que os atributos presentes na tabela base são, em sua maioria, texto, e, das métricas utilizadas, a similaridade de cosseno é a única que é compatível com dados textuais, as outras métricas são mais apropriadas para conjuntos numéricos. O segundo fator é o pré-processamento de dados, efetuado para todas as comparações (exceto no caso de ambas os atributos consistirem de números inteiros), que remove palavras que não estão presentes no vocabulário conhecido, resultando em conjuntos menores para comparação (ou por vezes removendo completamente os dados de um atributo), a maior parte das remoções ocorre com números grandes (acima de 10000).

Por consequência disso, para efetuar as análises necessárias, optamos por utilizar apenas a parte da estrutura referente à similaridade de cosseno neste trabalho. Esse problema não é causado pelo uso da estrutura - pois ela suporta tantas métricas de comparação quanto forem necessárias, basta adicionar o respectivo campo - mas pela tabela base utilizada. O uso da estrutura proposta permite mais facilmente comparar diferentes métricas de similaridade e analisar como os resultados foram obtidos para cada métrica.

Nome tabela	Cos	σ_{Cos}	$Jacc$	σ_{Jacc}	$Cscr$	σ_{Cscr}	CDF	σ_{cdf}	Média
regiao	0,674	0,000	0,002	0,002	0,002	0,002	0,500	0,000	0,294
fonte	0,639	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,160
ies_ens_superior	0,621	0,002	0,117	0,007	0,133	0,011	0,250	0,000	0,280
prouni	0,611	0,002	0,117	0,007	0,136	0,015	0,125	0,000	0,247
fies	0,625	0,001	0,113	0,006	0,132	0,011	0,251	0,002	0,280
curso_ens_superior	0,624	0,004	0,113	0,003	0,227	0,051	0,350	0,053	0,329
aluno_ens_superior	0,620	0,002	0,000	0,000	0,113	0,040	0,313	0,106	0,261
local_ens_superior	0,601	0,003	0,118	0,007	0,187	0,010	0,250	0,000	0,289
sim	0,661	0,002	0,112	0,003	0,307	0,007	0,388	0,040	0,367
institution_private_ag	0,731	0,003	0,013	0,017	0,050	0,051	0,000	0,000	0,198
institution_fies_ag	0,681	0,002	0,186	0,010	0,222	0,024	0,200	0,000	0,322
terras_indigenas*	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000
institution_ag	0,626	0,000	0,152	0,003	0,171	0,004	0,167	0,000	0,279
institution_prouni_ag	0,666	0,002	0,229	0,012	0,270	0,030	0,250	0,000	0,354
indigenas_territ_ag	1,000	0,000	1,000	0,000	1,000	0,000	1,000	0,000	1,000

Tabela 4.1: Médias e desvio padrões das similaridades

Outra razão para esta escolha é que como o experimento envolvia aplicar técnicas de anonimização e verificar se a similaridade era reduzida. Neste quesito, apenas a similaridade de cosseno apresentou valores altos o suficiente para que a analise fosse feita quando a similaridade fosse reduzida.

Para fazer a análise dos resultados, escolhemos estas quinze tabelas com similaridade alta e as separamos em quatro grupos de acordo com a diferença no número de colunas da tabela com a tabela base. Assim, obtivemos os seguintes grupos de tabelas: as que têm menos colunas que a tabela base, aquelas que têm aproximadamente o mesmo número de colunas, as que têm mais colunas e as que têm significativamente mais colunas.

A partir dessa divisão, escolhemos uma tabela de cada grupo para analisar com o auxílio da estrutura. Podemos, então, apresentar a estrutura referente à primeira amostra para exemplificar nosso processo de análise, alertamos, porém, que pode haver pequenas discrepâncias entre os valores nas tabelas e os presentes na estrutura, devido ao fato de que as tabelas contêm a agregação das dez amostras.

A listagem 4.1 expõe a parte da estrutura referente à similaridade de cosseno entre as duas tabelas. As linhas 2 a 4 contém informações sobre a métrica de comparação. As linhas 6 a 10 correspondem a comparação entre as colunas "ano" e "ano_censo". as linhas 11 a 15 contém informações sobre a comparação entre "tamanho_superficie" e "cod_ies". E as linhas 17 a 20 contém informações referentes a comparação entre as colunas "nome_municipio" e "nome_ies".

Uma análise inicial mostra que para as três colunas que foram consideradas mais similares, não há uma disparidade muito grande entre os valores calculados e que todas as colunas deveriam conter informações similares.

Uma análise um pouco mais detalhada mostra que, o primeiro par de colunas realmente é similar, pois ambas as colunas contêm informações sobre ano, por exemplo, no entanto, a similaridade não chega a 1.0 porque as colunas contêm anos únicos. Já o segundo par de colunas se revela um falso positivo, pois apesar de o conteúdo das colunas ser similar (ambas contém uma série de números, com grande parte estando entre 0 e 100), a informação presente é completamente diferente, enquanto uma contém a área de determinadas comunidades indígenas, a outra possui códigos que se referem a instituições de ensino. Por fim, o último par comparado também é um falso positivo, pois apesar da alta similaridade, a informação presente é diferente,

uma contém nomes de municípios, enquanto a outra contém nomes de instituições de ensino. A confusão ocorre porque muitas instituições de ensino municipais são identificadas como "Escola municipal de <nome do município>".


```
{
  "Metrica":"Distancia de cosseno",
  "explicacao":"https://en.wikipedia.org/wiki/Cosine_similarity",
  "Valor":0.729,
  "Comparação colunas":[
     {
        "Colunas comparadas":["ano", "ano_censo"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.620
     },{
        "Colunas comparadas":["tamanho_superficie", "cod_ies"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.751
     },{
        "Colunas comparadas":["nome_municipio", "nome_ies"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.817
     }
  ]
}
```


Listagem 4.1: Similaridade de cosseno para institution_private_ag

Do conjunto de tabelas com quantidade de colunas similares a tabela base, escolhemos a tabela institution_ag para mostrar a análise da estrutura. A estrutura da comparação feita pela similaridade de cosseno pode ser observada na listagem 4.2.

As linhas 2 a 4 contém informações sobre a métrica de comparação. As linhas 6 a 10 correspondem a comparação entre as colunas "sigla_uf" e "sigla_uf_ies". as linhas 11 a 15 contém informações sobre a comparação entre "ano" e "ano_cemso". As linhas 16 a 20 contém informações referentes a comparação entre as colunas "nome_terra" e "sigla_ies". As linhas 20 a 24 contém informações referentes as colunas "tamanho_superficie" e "cod_ies". As linhas 25 a 29 contém informações referentes as colunas "nome_municipio" e "nome_ies". E as colunas 30 a 34 contém informações a respeito das colunas "tipo_modalidade" e "region".

É possível observar que, neste caso, obtivemos uma mesclagem maior entre similaridades altas e baixas resultando em uma similaridade de tabelas 0.6. Observa-se também que os pares de colunas com similaridade baixa (abaixo de 0.5) é intuitivo pelos nomes das colunas que elas contém informações diferentes, a análise dos dados apenas confirmou este fato. Já para as que obtiveram similaridade acima de 0.5, os pares de sigla e ano possuem realmente a mesma informação, siglas de estados e anos respectivamente, com a diferença sendo siglas de estados (ou anos) em particular que uma tabela tem que a outra não.

Por outro lado, os pares com tamanho_superfície e nome_município apresentam os mesmos falso positivos que a tabela anterior. Esse resultado se deve, provavelmente ao fato de que as duas tabelas comparadas com "terras_indigenas" contêm informações sobre instituições de ensino, a diferença é que "institution_private_ag" refere-se a instituições privadas de ensino superior, enquanto "institution_ag" faz alusão a instituições de ensino superiorl.

```
{
   "Metrica":"Distancia de cosseno",
   "Explicacao":"https://en.wikipedia.org/wiki/Cosine_similarity",
   "Valor":0.626,
   "Comparação colunas":[{
        "Colunas comparadas":["sigla_uf","sigla_uf_ies"],
        "N Linhas base":738,
        "N Linhas comparada":103,
        "Valor":0.888
     }, {
        "Colunas comparadas":["ano", "ano_censo"],
        "N Linhas base":738,
        "N Linhas comparada":103,
        "Valor":0.617
     }, {
        "Colunas comparadas":["nome_terra","sigla_ies"],
        "N Linhas base":738,
        "N Linhas comparada":103,
        "Valor":0.403
     }, {
        "Colunas comparadas":["tamanho_superficie","cod_ies" ],
        "N Linhas base":738,
        "N Linhas comparada":103,
        "Valor":0.789
     }, {
        "Colunas comparadas":["nome_municipio","nome_ies"],
        "N Linhas base":738,
        "N Linhas comparada":103,
        "Valor":0.735
     }, {
        "Colunas comparadas":["tipo_modalidade", "region"],
        "N Linhas base":738,
        "N Linhas comparada":103,
        "Valor":0.322
     }]
}
```

Listagem 4.2: Similaridade de cosseno para institution_ag

Para apresentar a análise das tabelas com mais colunas que a tabela base, escolhemos tabela “fies”. A seção da estrutura referente à similaridade de cosseno entre as tabelas pode ser observada na listagem 4.3.

As linhas 2 a 4 contém informações sobre a métrica de comparação. As linhas 5 a 9 correspondem a comparação entre as colunas “fase_procedimento” e “curso”. as linhas 10 a 14 contém informações sobre a comparação entre “sigla_uf” e “sigla_uf”. As linhas 15 a 18 contém informações referentes a comparação entre as colunas “ano” e “ano_processo”. As linhas 19 a 23 contém informações referentes as colunas “nome_terra” e “campus”. As linhas 24 a 28 contém informações referentes as colunas “tamanho_superficie” e ‘mes_processo”. As colunas 29 a 33 contém informações a respeito das colunas “nome_etnia” e “nome_ies_fies_ext_aluno”. As colunas 34 a 38 contém informações referentes as colunas “nome_municipio” e “nome_municipio”. E a s linhas 39 a 43 contém informações referentes as colunas “tipo_modalidade” e “nome_ies”.

É perceptível que, neste caso, apesar da similaridade não ser muito alta, nas similaridades entre colunas há uma distribuição boa de atributos com similaridades altas e baixas, isso faz com que a média geral seja mais baixa.

```
{  "Metrica":"Distancia de cosseno",
   "explicacao":"https://en.wikipedia.org/wiki/Cosine_similarity",
   "Valor":0.625,
   "Comparação colunas":[{
        "Colunas comparadas":["fase_procedimento", "curso" ],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.386
     }, {
        "Colunas comparadas":["sigla_uf","sigla_uf"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.827
     }, {
        "Colunas comparadas":["ano","ano_processo" ],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.479
     }, {
        "Colunas comparadas":["nome_terra","campus"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.746
     }, {
        "Colunas comparadas":["tamanho_superficie","mes_processo"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.847
     }, {
        "Colunas comparadas":["nome_etnia","nome_ies_fies_ext_aluno" ],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.420
     }, {
        "Colunas comparadas":["nome_municipio","nome_municipio"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.924
     }, {
        "Colunas comparadas":["tipo_modalidade","nome_ies"],
        "N Linhas base":738,
        "N Linhas comparada":1500,
        "Valor":0.375
     }]
},
```


Listagem 4.3: Similaridade de cosseno para fies

Dentre os pares de colunas pouco similares (abaixo de 0.5), apenas o par relativo a anos é um falso positivo evidente, causado pela discrepância entre os anos presentes nas tabelas, o restante dos pares pouco similares é intuitivo a partir do nome (e confirmado com os dados) que os atributos se referem a informações diferentes.

Já para os pares de atributos considerados similares, com exceção dos pares de "siglas_uf" e "nome_município", que realmente contêm a mesma informação, os outros são falsos positivos. O problema se repete no par "tamanho_superfície" é causado por razões similares às

outras instâncias e “nome_terra” com “campus” é também por razões similares ao erro entre “nome_municipio” e “nome_ies” apesar dos 4 atributos conterem informações diferentes.

Por fim, para representar a análise das tabelas que contêm significativamente mais colunas que a tabela base, escolhemos a tabela “sim”, a respectiva seção da estrutura pode ser observada na listagem 4.4. Como foi o caso na comparação anterior, a similaridade geral não é muito alta devido ao fato de que, durante o cálculo da média, os pares com similaridade alta são compensados pelos pares de atributos com similaridades baixas.

As linhas 2 a 4 contém informações sobre a métrica de comparação. As linhas 5 a 9 correspondem a comparação entre as colunas “fase_procedimento” e “defz_fonte”. as linhas 10 a 14 contém informações sobre a comparação entre “sigla_uf” e “res_SIGLA_UF”. As linhas 15 a 18 contém informações referentes a comparação entre as colunas “ano” e “ano_obito”. As linhas 19 a 23 contém informações referentes as colunas “nome_terra” e “res_MUNNOME”. As linhas 24 a 28 contém informações referentes as colunas “tamanho_superficie” e “idade_obito_anos”. As colunas 29 a 33 contém informações a respeito das colunas “nome_etnia” e “res_MUNNOMEX”. As colunas 34 a 38 contém informações referentes as colunas “nome_municipio” e “ocor_MUNNOME”. E a s linhas 39 a 43 contém informações referentes as colunas “tipo_modalidade” e “def_gravidez”.

```
{ "Metrica":"Distancia de cosseno",
  "explicacao":"https://en.wikipedia.org/wiki/Cosine_similarity",
  "Valor":0.661,
  "Comparação colunas":[{
      "Colunas comparadas":[ "fase_procedimento","def_fonte"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.449
    }, {
      "Colunas comparadas":["sigla_uf","res_SIGLA_UF"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.853
    }, {
      "Colunas comparadas":[ "ano", "ano_obito"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.361
    }, {
      "Colunas comparadas":[ "nome_terra","res_MUNNOME"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.880
    }, {
      "Colunas comparadas":["tamanho_superficie","idade_obito_anos"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.936
    }, {
      "Colunas comparadas":["nome_etnia","res_MUNNOMEX"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.426
    }, {
      "Colunas comparadas":["nome_municipio","ocor_MUNNOME"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.961
    }, {
      "Colunas comparadas":["tipo_modalidade","def_gravidez"],
      "N Linhas base":738,
      "N Linhas comparada":1500,
      "Valor":0.425
    }]
},
```


Listagem 4.4: Similaridade de cosseno para sim

Analisando a estrutura e os dados, vemos que, dos pares com similaridade baixa, apenas o par referente a "anos" poderia ser um falso negativo. No caso do atributo "ano", em todos os cenários em que o consideramos como um falso negativo, foi apenas no quesito de ambos os atributos se referirem a anos, dependendo de informações adjacentes (que só seriam descobertas em análise após a busca) a similaridade baixa é justificada, como foi o caso nesta comparação, em que a tabela base tem os anos em que um processo judicial começou, a tabela comparada "sim" contém informações sobre o ano de óbito de indivíduos.

Para os pares que obtiveram similaridade alta, os referentes a sigla de UFs, nome de municípios e nome de terra realmente são pares similares, com "tamanho_superfície" sendo um falso positivo assim como nos outros casos.

Uma vez concluída esta análise, aplicamos a técnica de supressão em quatro dos atributos da tabela base e refizemos a busca por similares. Os atributos escolhidos foram "fase_procedimento", "nome_terra", "nome_etnia" e "nome_município", estes atributos foram escolhidos porque consideramos que essa divisão permitiria verificar mais facilmente o efeito da anonimização na busca por similares efetuada.

A análise inicial do resultado desta segunda busca por similares mostrou que apenas cinco tabelas obtiveram uma similaridade acima de 0.5, sem que nenhuma chegasse 0.6.

Como a supressão removeu todos os valores das colunas suprimidas, o tempo para fazer comparações diminuiu, como era o esperado, já que colunas vazias automaticamente retornam um código de erro informando que não é possível efetuar a comparação, este código de erro é tratado como uma similaridade 0 entre colunas em todos os pontos em que é relevante para algum cálculo, ele é mantido como um código diferente para que, na estrutura de explicação, seja possível ver que a comparação não ocorreu.

As tabelas que ficaram com a similaridade acima de 0.5 são aquelas em que a similaridade vem de falsos positivos (com exceção da tabela base e da tabela agregada a partir da tabela base), então, efetivamente, após o processo de anonimização, não restou nenhuma tabela similar à tabela de base. As similaridades calculadas para as tabelas após a anonimização da tabela base, para as tabelas com similaridade acima de 0.5 podem ser encontradas na tabela 4.2.

Nome tabela	Cos	σ_{Cos}	$Jacc$	σ_{Jacc}	$Cscore$	σ_{Cscore}	CDF	σ_{CDF}	Média
regiao	0,566	0,000	0,001	0,003	0,001	0,003	0,500	0,000	0,267
fonte	0,585	0,000	0,005	0,010	0,005	0,010	0,021	0,043	0,154
institution_pri_ag	0,591	0,001	0,002	0,005	0,010	0,020	0,000	0,000	0,151
terras_indigenas*	0,500	0,000	0,500	0,000	0,500	0,000	0,500	0,000	0,500
indigenas_terr_ag	0,500	0,000	0,500	0,000	0,500	0,000	0,500	0,000	0,500

Tabela 4.2: Médias e desvio padrões das similaridades após anonimização

Para as tabelas "região", "fonte" e "institution_private_ag" (na tabela intitution_pri_ag) a similaridade vem, principalmente, de um falso positivo com o atributo "tamanho_superficie". As únicas tabelas que possuem um valor de similaridade realmente relevante são a tabela base e a tabela construída a partir da tabela base. Mas mesmo elas têm a similaridade alta de seus atributos drasticamente reduzida quando é calculada a similaridade entre tabelas devido aos atributos anonimizados.

4.3 CONSIDERAÇÕES

O resultado dos experimentos mostra que apenas a aplicação de técnicas simples já é o suficiente para drasticamente alterar a similaridade da tabela anonimizada com as tabelas encontradas durante o processo de busca por similares.

Isso reforça a noção que anonimizar os dados de uma tabela afeta a busca por similares – e por consequência ataques de ligação – de forma que tabelas que eram consideradas similares possivelmente não sejam encontradas.

A explicação da busca por similares ajudou a fazer a análise dos resultados, isso possibilitou o acesso às informações de que precisávamos. A combinação da estrutura de explicação com uma interface visual para controlar melhor as informações exibidas pela estrutura

e o acesso ao banco para conferir se os resultados encontrados estavam corretos facilitou o processo de análise dos resultados, permitindo que trabalhássemos mais rapidamente.

Tendo em vista como a estruturação da explicação facilitou a analise do resultado da busca por similares é razoável assumir que essa facilidade se estenderia para a analise de riscos de anonimização – principalmente para encontrar quase-identificadores –, permitindo que sejam tomadas decisões melhores de como e quais atributos devem ser anonimizados.

Reconhecemos que o modo como definimos quais colunas serão consideradas similares para fazer o cálculo de similaridade entre tabelas (fazer pareamento maximal de um grafo contendo todas as comparações entre colunas) apresenta uma falha. Essa estratégia favorece tabelas com menos colunas.

Outro problema em nossa implementação é que não tomamos nenhum cuidado quando ha uma diferença grande entre a quantidade de elementos de cada tabela, o que pode enviesar o resultado obtido. Neste aspecto pode ser considerado como uma positivo que fomos forçados a trabalhar com amostras dos dados, de forma que não isso não causou problemas para nós.

Apesar da estrutura ter atendido às expectativas de facilitar a análise do resultado da busca por similares, ainda existem melhorias que podem ser feitas. Decidimos utilizar a linguagem **python** para fazer nosso experimento para servir como uma prova de conceito que a estrutura é fácil de montar e utilizar. Agora que sabemos que nossa ideia seria interessante fazer um estudo pra ver quanto é possível otimizar a geração da estrutura quando implementada em uma linguagem de programação de mais baixo nível, como C ou C++.

Um experimento a mais que teria sido interessante de executar, porém não seria possível executar em tempo hábil, seria fazer o experimento utilizando a base de dados completa em vez de amostras.

5 CONCLUSÕES

Nesta dissertação, apresentamos nossa proposta de como construir explicações estruturadas de forma a prover boas explicações de uma busca por similares para auxiliar o processo de análise de riscos da anonimização de dados.

Nossos estudos começaram na área de integração de dados abertos, porém logo foram redirecionados para a área de busca por similares em dados abertos. Esse redirecionamento ocorreu devido a nossa percepção de uma aparente dualidade entre busca por similares e anonimização de dados. A partir disso, estudamos diversos trabalhos a respeito de anonimização e comprovamos que, de fato, essa relação existia (Merener, 2012; Nargesian et al., 2018; Ozalp et al., 2016).

A existência dessa relação dual entre anonimização e busca por similares evidencia, de um lado, que a anonimização de dados prejudica a busca por similares porque remove informações dos atributos que poderiam ser utilizadas para calcular similaridade, por outro lado, a busca por similares pode ser utilizada para encontrar tabelas que contenham quase-identificadores de uma tabela anonimizada, facilitando o trabalho de um potencial atacante de encontrar estes quase-identificadores. Percebemos também que é possível utilizar esta relação de forma benéfica de forma que os resultados de uma busca por similares contribuam para a qualidade da anonimização.

Se a busca por similares for utilizada como parte do processo de anonimização, em particular na análise de risco para decidir como e quais atributos devem ser anonimizados, é possível melhorar a qualidade da análise de riscos e, consequentemente, da anonimização. Porém, apenas utilizar a busca não é o suficiente, para melhorar ainda mais a análise (e facilitar o processo de análise em si) percebemos que seria necessário explicar como o resultado da busca por similares foi obtido.

Após estudar alguns trabalhos na área de explicações de resposta (Wang et al., 2018; Bohlender e Köhl, 2019), percebemos que apesar de haver um certo consenso sobre os fatores que compõem uma boa explicação, não existem métricas para medir a qualidade de uma explicação. Além disso, a qualidade de uma explicação é muito relativa ao grupo alvo para o qual a explicação é dada.

A partir dessas percepções, propusemos um modelo para construir explicações em uma estrutura padronizada, de forma que seja fácil de manipular o modo como a explicação é apresentada. Essa mesma estrutura permite prover explicações de qualidade para diversos grupo-alvos diferentes usando uma interface simples, dado que o tipo de estrutura proposta já é muito utilizado e possui diversos interpretadores disponíveis.

Poder controlar como a explicação é apresentada facilita a utilização de fatores que fazem parte de uma explicação boa, por exemplo, é possível atiçar a curiosidade de quem recebe a explicação ou apresentar a explicação de maneiras diferentes para grupos diferentes. Ao mesmo tempo, é minimizada a ocorrência de fatores que podem pior a qualidade da explicação, como sobrecarga de informações.

Para testar nossa hipótese de que a busca por similares explicada permite fazer uma análise de riscos melhor antes de anonimizar dados, realizamos uma avaliação experimental. Os experimentos mostraram que mesmo as técnicas mais simples de anonimização foram o suficiente para alterar significativamente a similaridade calculada durante uma busca por similares, o que reforça a ideia de que anonimizar dados afeta de forma negativa a busca por similares.

Concluímos a partir dos experimentos que estruturar a explicação, e uma interface simples para apresentar a explicação facilitam a analise dos resultados da busca por similares,

e que portanto seria razoável estender esta facilidade de analise para a analise de riscos de anonimização como descrita no fluxo da seção 3.4.

Em suma, nosso trabalho apresenta uma contribuição para utilizar a busca por similares de modo a auxiliar a análise de riscos feita durante a anonimização, incluindo um fluxo de trabalho a ser seguido para tal. Além disso, outra contribuição deste estudo foi a elaboração de um modelo de construção de explicações que permite a construção de explicações melhores.

De maneira geral, os resultados obtidos sugerem que nossa abordagem é sustentável e aponta estudos posteriores que contribuiriam ainda mais para a área e preencheriam as lacunas naturais de uma dissertação de mestrado.

Em pequena escala melhorias poderiam ser feitas a partir deste trabalho, por exemplo desenvolver uma implementação mais eficiente do programa que monta a estrutura ou adaptar uma implementação eficiente já existente para utilizar a estrutura; realizar os experimentos em uma base de dados maior, com as tabelas completas ao invés de amostras.

Em maior escala trabalhos futuros deveriam focar em estudar de forma mais aprofundada como diferentes técnicas de anonimização – k-anonimidade, l-diversidade, etc – afetam a similaridade encontrada durante a busca por similaridades.

Outro ponto de interesse para trabalhos futuros é estudar a melhor forma de apresentar a explicação da busca por similares, tanto para pesquisadores que trabalham na área, quanto para leigos que tem curiosidade de entender o processo de busca.

Por fim, um ultimo foco que pode ser dado para trabalhos futuros é aplicar a ideia de estruturar a explicação em outras áreas de pesquisa além da busca por similares. Planejar como a explicação seria estruturada e organizada é uma boa maneira de definir quais são os componentes importantes que uma explicação precisa ter para adequadamente explicar um processo.

REFERÊNCIAS

Berlin, J. e Motro, A. (2002). Database schema matching using machine learning with feature selection. Em *International Conference on Advanced Information Systems Engineering*, páginas 452–466. Springer.

Bogatu, A., Fernandes, A. A., Paton, N. W. e Konstantinou, N. (2020). Dataset discovery in data lakes. Em *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, páginas 709–720. IEEE.

Bohlender, D. e Köhl, M. A. (2019). Towards a characterization of explainable systems. *arXiv preprint arXiv:1902.03096*.

Brickell, J. e Shmatikov, V. (2008). The cost of privacy: destruction of data-mining utility in anonymized data publishing. Em *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 70–78.

Došilović, F. K., Brčić, M. e Hlupić, N. (2018). Explainable artificial intelligence: A survey. Em *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, páginas 0210–0215. IEEE.

Hoffman, R. R., Mueller, S. T., Klein, G. e Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Holzinger, A., Carrington, A. e Müller, H. (2020). Measuring the quality of explanations: the system causability scale (scs). *KI-Künstliche Intelligenz*, páginas 1–6.

Li, N., Li, T. e Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. Em *2007 IEEE 23rd International Conference on Data Engineering*, páginas 106–115. IEEE.

Machanavajjhala, A., Kifer, D., Gehrke, J. e Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es.

Merener, M. M. (2012). Theoretical results on de-anonymization via linkage attacks. *Transactions on Data Privacy*, 5(2):377–402.

Miller, R. J. (2018). Open data integration. *Proceedings of the VLDB Endowment*, 11(12):2130–2139.

Murthy, S., Bakar, A. A., Rahim, F. A. e Ramli, R. (2019). A comparative study of data anonymization techniques. Em *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, páginas 306–309. IEEE.

Nargesian, F., Zhu, E., Pu, K. Q. e Miller, R. J. (2018). Table union search on open data. *Proceedings of the VLDB Endowment*, 11(7):813–825.

Ozalp, I., Gursoy, M. E., Nergiz, M. E. e Saygin, Y. (2016). Privacy-preserving publishing of hierarchical data. *ACM Transactions on Privacy and Security (TOPS)*, 19(3):1–29.

Rosenfeld, A. (2021). Better metrics for evaluating explainable artificial intelligence. Em *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, páginas 45–50.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

Wang, X., Haas, L. e Meliou, A. (2018). Explaining data integration. *Data Engineering Bulletin*, 41(2).

Xiao, X. e Tao, Y. (2007). M-invariance: towards privacy preserving re-publication of dynamic datasets. Em *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, páginas 689–700.

Zhu, E., Deng, D., Nargesian, F. e Miller, R. J. (2019). Josie: Overlap set similarity search for finding joinable tables in data lakes. Em *Proceedings of the 2019 International Conference on Management of Data*, páginas 847–864.