Instituto Federal da Paraíba

Mestrado Profissional em Tecnologia da Informação

Relatório Técnico

João Pessoa, 13 de abril de 2020.

Integração de Dados no Contexto de Acidentes de Trânsito: fundamentos e avaliação de algoritmos de *clusterização* no apoio à resolução de entidades

Maria Aparecida da Silva Santiago, Damires Yluska de Souza Fernandes[1]

[1]Unidade Acadêmica de Informática

Resumo

Em ambientes de compartilhamento de dados, muitas aplicações podem requisitar acesso integrado a diversas fontes de dados que são autônomas e heterogêneas. A Integração de Dados é realizada em etapas, sendo uma delas a Resolução de Entidades. Nesse panorama, este relatório apresenta os resultados obtidos em um estudo que utilizou métodos de clusterização aplicados ao problema de Resolução de Entidades no contexto de dados de acidentes de trânsito. Na avaliação realizada, foi aplicado um processo de extração, transformação e carga dos dados juntamente com a tarefa de clusterização, focados na Resolução de Entidades. Os algoritmos avaliados foram o K-Means e o K-Medoids. A análise preliminar dos resultados da aplicação dos algoritmos ao problema em questão indica que o K-Means apresentou resultados ligeiramente melhores que o K-Medoids, em relação à quantidade de instâncias corretamente identificadas, de acordo com avaliação dos Especialistas do Domínio.

1. Introdução

Nos últimos anos a Organização Mundial da Saúde tem alertado para o crescimento contínuo de mortes por acidentes de trânsito. Em alguns países, o problema já é tratado como epidemia, de acordo com o *Global status report on road safety 2018* (OMS, 2018). No Brasil, os acidentes de trânsito representam um dos principais problemas de saúde pública presentes do país devido a sua elevada taxa de morbimortalidade, sobrecarga no sistema de saúde e repercussão social (LIMA et al., 2019). Na cidade de João Pessoa, foram registrados 6.917 acidentes de trânsito de janeiro a outubro de 2019. Esses números são baseados nos registros policiais, nas informações dos Serviços de Atendimento Móvel de Urgência (SAMU)[1] e em dados das bases dos hospitais especializados em traumas. Contudo, estima-se que os números reais são ainda maiores, quando considerados os acidentes com lesões de baixa gravidade onde a vítima deixa o local do acidente e não gera registro.

Na cidade de João Pessoa, os dados de acidentes de trânsito são gerenciados pelo órgão de trânsito da cidade. O papel do órgão é coletar os dados providos pelas diferentes fontes (SAMU, hospitais de trauma e polícia civil) para então realizar análises. Essas análises servem como indicadores socioeconômicos para os gestores de trânsito avaliarem estratégias para mitigação de acidentes, que, por sua vez, podem reduzir os números de mortes ou traumas irrecuperáveis.

Os dados dos hospitais fornecem informações sobre as lesões sofridas pelas vítimas como também dados demográficos sobre o acidente e a vítima. Os dados do SAMU fornecem maiores detalhes sobre as condições do acidente como: envolvidos, condição da vítima, local e hospital de destino para o qual a vítima foi encaminhada. Os dados da polícia fornecem detalhes sobre as características do acidente como dados dos veículos e do condutor. Dessa maneira, os registros são complementares e representam uma rica fonte de informação para o órgão de segurança viária. No entanto, realizar a integração dos dados entre as diferentes fontes de dados de maneira não automatizada é um processo demorado e sujeito a elevadas quantidades de falhas (KAMALUDDIN et al., 2018).

A Integração de Dados tem o objetivo de fornecer acesso unificado a dados residentes em múltiplas e autônomas fontes de dados (DONG e SRIVASTAVA, 2015). Embora a definição seja aparentemente simples, alcançar esse objetivo não tem sido fácil mesmo para um pequeno número de fontes e mesmo para dados estruturados (DOAN et al. 2012), devido principalmente às questões de heterogeneidade semântica. De acordo com Dong e Srivastava (2015) a Integração de Dados inclui três etapas principais: Mapeamento de Esquemas, Resolução de Entidades e Fusão de Dados.

Considerando as ambiguidades semânticas e/ou sintáticas, o Mapeamento de Esquemas consiste em mapear esquemas diferentes de fontes que se referem ao mesmo domínio de dados. Isso pode ser realizado dos esquemas das fontes para um esquema global ou entre os esquemas (DONG e SRIVASTAVA, 2015). A segunda etapa na integração de dados é a Resolução de Entidades que aborda o desafio de detectar a ambiguidade de representação de instâncias e visa entender quais registros representam a mesma entidade no mundo real. A terceira etapa necessária está

1 http://saude.gov.br/saude-de-a-z/servico-de-atendimento-movel-de-urgencia-samu-192

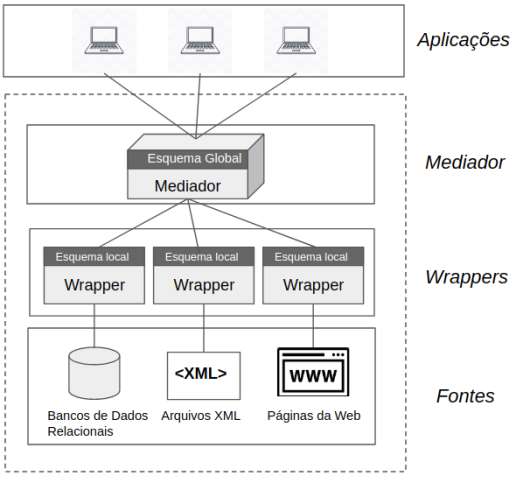
associada à necessidade de verificação de conflitos entre os dados dos registros e seus valores (DONG et al., 2014). Este problema é definido como combinação de dados (*data merge*) ou fusão de dados (*data fusion*). Assim, após a resolução de entidades, os registros que representam a mesma entidade do mundo real passam pelo processo de fusão e são unificados em uma única representação.

Este relatório apresenta resultados de um estudo sobre Resolução de Entidades no domínio de dados de acidentes de trânsito. Para tanto foram utilizados dados reais de acidentes dos anos de 2010 a 2019 da cidade de João Pessoa. O objetivo principal deste trabalho foi avaliar a viabilidade do uso de algoritmos de clusterização de dados, aplicados ao problema de Resolução de Entidades para ajudar a automatizar o processo de Integração de Dados utilizando uma abordagem materializada. Profissionais da Superintendência Executiva de Mobilidade Urbana de João Pessoa (SEMOB-JP) atuaram como Especialistas do Domínio de trânsito para auxílio às tarefas citadas.

2. Sistemas de Integração de Dados

A Integração de Dados é realizada a partir de técnicas para combinar dados de fontes heterogêneas de maneira que informações úteis possam ser utilizadas de modo complementar e conjunto. O objetivo de um sistema de integração de dados é oferecer aos usuários uma interface uniforme de acesso a diferentes fontes de dados, de forma que os usuários definam consultas especificando “o que” se deseja saber e o sistema determine “onde e como” a informação pode ser encontrada e, em seguida, apresente as respostas já integradas para as consultas do usuário (LÓSCIO e SALGADO, 2013). Algumas características relacionadas aos dados como sua localização e modelo de dados empregado são fatores que contribuem para que o problema de integrar dados seja um desafio. Quanto à localização, os dados podem estar fisicamente distribuídos em diferentes locais de armazenamento, por exemplo em vários servidores distribuídos pelo mundo. Quanto ao modelo, os dados podem estar estruturados de diferentes maneiras em nível de esquema e de dados. Dessa maneira, a integração de dados é realizada com o objetivo de que seja possível utilizar dados a partir de suas fontes, independentemente do lugar onde estão armazenados e de seu nível de estruturação. A Figura 1 apresenta um modelo de arquitetura de um Sistema de Integração de Dados.

Figura 1: Sistemas de Integração de Dados.



Fonte: Adaptado de Katsis e Papakonstantinou (2018).

De acordo com a Figura 1, as fontes (s*ources*) armazenam dados que podem estar representados em diversos formatos como tabelas de um banco de dados relacional, arquivos XML[2] ou páginas escritas em HTML[3]. Os *wrappers* são responsáveis por resolver o problema de heterogeneidade nos dados das fontes, transformando os modelos de dados individuais de cada fonte em um modelo de dados comum ao sistema de integração de dados. O esquema das fontes de origem é chamado de Esquema Local (*Local Schema*). O papel do mediador (*mediator*) é fornecer um Esquema Global (*Global Schema*) para que haja uma visão unificada das fontes para as aplicações (*applications*), tornando transparente para elas a conversão das consultas para um esquema local específico de cada fonte.

As subseções seguintes ampliam alguns conceitos e apresentam abordagens existentes para integração de dados.

2.1 Heterogeneidade de Dados

A heterogeneidade em fontes de dados está ligada, principalmente, ao uso de diferentes modelos de dados e podem ser classificadas como sintática, semântica ou estrutural (LÓSCIO e SALGADO, 2013). A heterogeneidade sintática está relacionada ao formato de representação de um dado pertencente a um esquema. Por exemplo, considerando uma empresa que mantém dados de seus clientes em tabelas de bancos de dados relacionais distribuídos entre as suas unidades, podem existir os valores ‘0’ e ‘1’ ou ‘M’ e ‘F’ para representação do atributo sexo de um cliente. Isto é, o mesmo atributo é representado com sintaxes diferentes em fontes diversas. Essas divergências são causadas por diferentes representações para um mesmo dado, quando o dado está presente em mais de uma fonte.

Heterogeneidade semântica diz respeito ao significado dos metadados em um dado domínio ou contexto. Conflitos semânticos podem ocorrer quando metadados são representados da mesma maneira (homófonas), mas possuem significados diferentes ou metadados cujas descrições possuem grafias diferentes, mas possuem o mesmo significado (homônimas). Diferentes interpretações sobre os dados causam heterogeneidade semântica nas fontes de dados. Por exemplo, supondo que existem duas fontes de dados A e B, que agregam dados sobre pacientes de um hospital, a fonte A poderia representar o registro da data de chegada do paciente ao hospital pelo metadado ‘Entrada’ enquanto, na fonte B, o mesmo registro é representado utilizando o metadado ‘Chegada’. Nesse caso particular, existe divergência semântica na representação do metadado.

A heterogeneidade estrutural está relacionada com a estrutura em que o formato é armazenado, ou seja, como existem esquemas de armazenamento de dados distintos, um dado pode estar representado em um esquema por uma estrutura, enquanto em outro esquema, o mesmo dado pode estar representado de uma maneira diferente. Por exemplo, considerando o modelo relacional, em um esquema de dados, o endereço de um cliente pode estar estruturado em formato de tabela enquanto, em outro esquema, o endereço pode estar armazenado como atributo simples (string) pertencente a uma tabela. Outro problema dessa natureza pode ser com relação aos formatos, por exemplo,

2 https://www.w3.org/XML/
3 https://www.w3.org/html/

uma fonte pode disponibilizar seus dados em formato JSON[4] enquanto outra fonte pode disponibilizar seus dados em formato CSV[5].

O problema da integração de dados está diretamente relacionado com essas questões de heterogeneidade.

2.2 Abordagens para Integração de Dados

Na construção de um sistema de integração de dados, uma importante decisão a ser tomada é a escolha da abordagem a ser utilizada. De acordo com Lóscio e Salgado (2013), as abordagens básicas a serem consideradas são a virtual e a materializada.

Na abordagem virtual, os dados são extraídos das fontes dinamicamente no momento em que eles são requisitados. Na abordagem materializada, os dados são extraídos previamente das fontes e armazenados em um repositório de forma que, quando uma consulta é realizada sobre os dados, a recuperação do dado é feita no repositório onde os dados estão armazenados e não diretamente nas fontes. Tanto a abordagem virtual quanto a materializada possuem vantagens e desvantagens. O Quadro 1 resume as principais vantagens e desvantagens de ambas abordagens.

Quadro 1. Comparação entre as abordagens virtual x materializada (LÓSCIO e SALGADO, 2013).

Abordagem	Vantagens	Desvantagens
Virtual	A garantia de que os dados estão sempre atualizados é maior pois são extraídos das fontes de origem sempre que são requisitados através de uma consulta.	A disponibilidade dos dados é completamente dependente da disponibilidade da fonte, se uma fonte estiver inacessível pode comprometer o sistema de integração de dados.
Materializada	Adequada quando o tempo de resposta à consulta deve ser mais importante para quem está requisitando do que o estado de atualização dos dados.	Podem ocorrer inconsistências nos dados do repositório em relação aos dados das fontes.

É importante destacar que não existe uma regra sobre quando deve ser utilizada uma abordagem ou outra, tudo vai depender das características da(s) aplicação(ões) que fará(ão) uso dos dados. Em alguns casos, ambas abordagens podem ser utilizadas de modo híbrido. Para implementação dessas abordagens, duas arquiteturas foram definidas e têm sido base para outras mais recentes, a saber: (i) arquitetura de Mediadores e (ii) *Data Warehouse*. Elas são detalhadas a seguir.

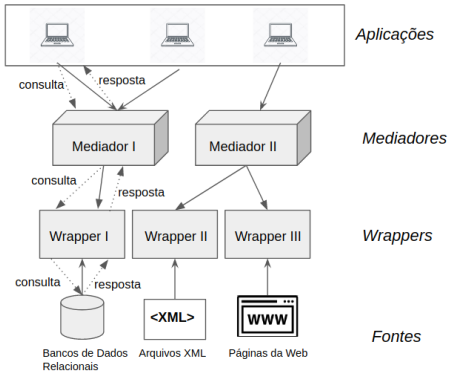
2.2.1 Arquitetura de Mediadores

Os mediadores são módulos de software que exploram o conhecimento

4 https://www.json.org/json-en.html
5 https://www.w3.org/TR/tabular-data-primer/

representado em um conjunto ou subconjunto de dados para gerar informações para aplicações residentes em uma outra camada (em geral, superior) (WIEDERHOLD, 1992). Arquiteturas baseadas em mediadores implementam a abordagem virtual para integração de dados. Nesse caso, as consultas são submetidas ao sistema de integração através do mediador, que é responsável por decompor as consultas em subconsultas a serem enviadas às fontes de dados (LÓSCIO e SALGADO, 2013). O papel principal do mediador no sistema de integração de dados é fornecer um nível de abstração em forma de esquema integrado entre as fontes de dados. Assim as consultas e atualizações são realizadas através do mediador, e o mediador faz suas traduções em forma de subconsultas diretamente para as respectivas fontes de dados. A Figura 2 ilustra a arquitetura de mediadores de acordo com a definição de Abiteboul (2000) conforme explicado.

Figura 2: Arquitetura de Mediadores.



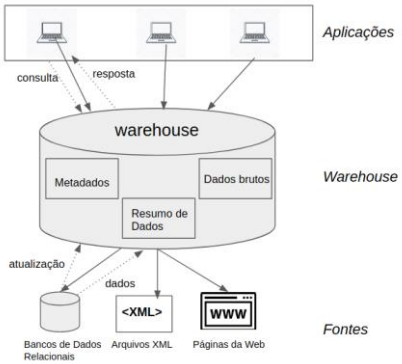
Fonte: Adaptado de Abiteboul, 2000.

A Arquitetura de Mediadores possui as mesmas vantagens e desvantagens destacadas na abordagem virtual.

2.2.2 Arquitetura de *Data Warehouse*

Assim como a arquitetura de mediadores é baseada na abordagem virtual, a arquitetura de *Data Warehouse* é uma forma de implementar a abordagem materializada. Nesse caso, os dados são recuperados, integrados e armazenados em um repositório de dados (LÓSCIO e SALGADO, 2013), que é referenciado como armazém de dados (*Data Warehouse*). A Figura 3 (ABITEBOUL, 2000) mostra a arquitetura de integração de dados com data warehouse.

Figura 3: Arquitetura de Data Warehouse**.**



Fonte: Adaptado de Abiteboul, 2000.

De acordo com essa arquitetura ilustrada através da Figura 3, os dados das fontes podem ser de diversos tipos, estruturas, formatos etc. O módulo servidor é responsável por sincronizar os dados entre o *warehouse* e as fontes. Esta abordagem possui as mesmas vantagens e desvantagens destacadas na abordagem materializada.

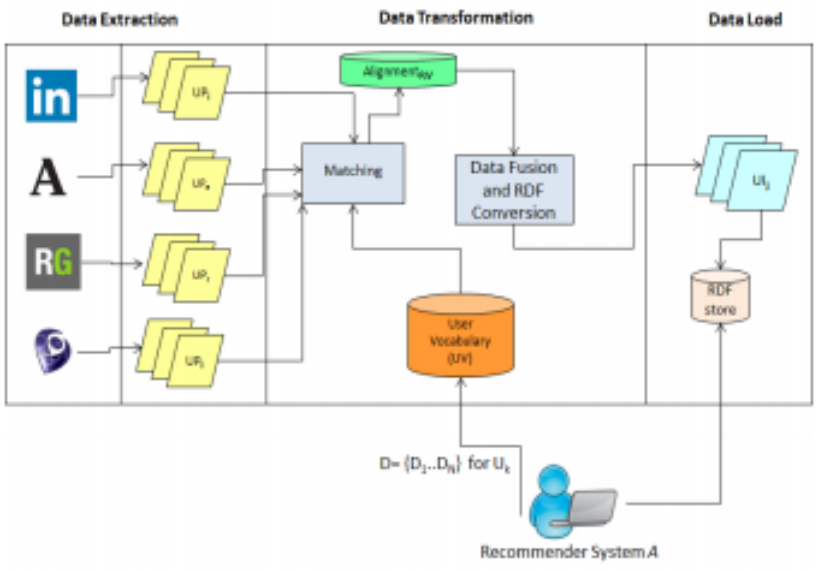
A Integração de Dados utilizando a arquitetura de *Data Warehouse/materializada* é normalmente realizada por meio de um processo de Extração-Transformação-Carga ou *Extract-Transform-Load* (ETL) (BANSAL e KAGEMANN, 2014). Refere-se a um processo que extrai dados de fontes diversas, transforma-os conforme um modelo que possa ser utilizado por aplicações e carrega os dados integrados em um repositório de destino final (BANSAL e KAGEMANN, 2014). As fases do processo ETL são descritas a seguir (ALENCAR et al., 2018; BANSAL e KAGEMAN, 2014):

- *Extração*: A primeira etapa do processo envolve a coleta e seleção dos dados nas fontes, que podem estar em diferentes níveis de estruturas (estruturado, não-estruturado ou semiestruturado), ou em diferentes formatos como CSV, XML, JSON, TXT, etc;

Transformação: Esta etapa do processo envolve tarefas como limpeza ou normalização dos dados, resolução de conflitos sintáticos, estruturais, semânticos e de dados e conversão dos dados de seu formato original para um modelo de destino integrado;
Carga: Essa é a fase do armazenamento dos dados em um repositório implementado por meio, por exemplo de um SGBD.

O processo de integração de dados por meio de ETL pode ser aplicado em diferentes ambientes de dados distribuídos. A Figura 4 ilustra resumidamente um exemplo de processo ETL para integração de dados, em ambiente Web.

Figura 4: Integração de Dados – abordagem materializada**.**



Fonte: Alencar et al., 2018.

Uma possível abordagem para o processo de Integração de Dados envolve as etapas de coleta dos dados das fontes, um processo de transformação, o armazenamento e a utilização final dos dados, por exemplo, para atividades de análises de dados.

No contexto de *Big Data*, a Integração de Dados (BDI - *Big Data Integration*) enfrenta ainda mais desafios por conta da própria natureza dos dados. De acordo com Dong e Srivastava (2015), primeiramente, o *volume* dos dados existentes nesse contexto é em grandezas de milhões de dados. O processo para integrar esses dados deve levar também em consideração a quantidade de dados a ser processados. Em segundo lugar, os dados são gerados e consumidos dinamicamente então a *velocidade* com que estes dados devem ser processados é outro fator imprescindível para BDI pois precisa acompanhar a necessidade das aplicações que utilizam tais dados em tempo real. A *variedade* entre os dados é outro fator importante, pois existem muitas formas tanto em estruturas quanto em conteúdos quando se trata de *Big Data* e seu contexto, muitas vezes, associado à Internet das Coisas (IoT). A *veracidade* dos dados também é um dos principais fatores discutidos pois, em BDI, análises sobre os dados são feitas em tempo real para produzir resultados dinamicamente. Se os dados utilizados para tais análises não forem confiáveis, o processo de BDI não irá agregar valor ao propósito ao qual se destina. A integração de dados é fundamental para que, de fato, as vantagens em *Big Data* possam ser obtidas.

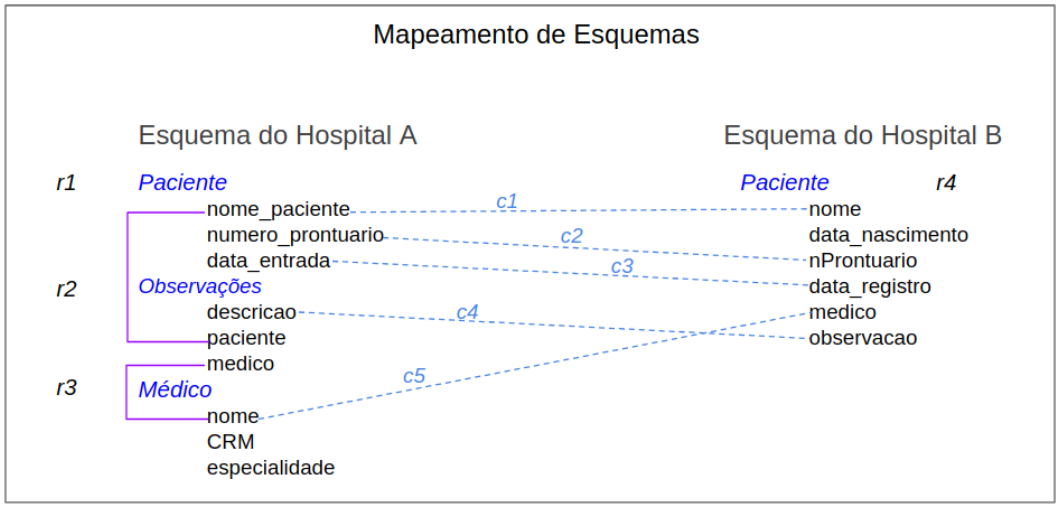
Assim como na integração de dados “tradicional”, a BDI também enfrenta desafios para tratamento de conflitos sintáticos, estruturais, semânticos e de dados (DONG e SRIVASTAVA, 2015).

Segundo Dong e Srivastava (2015), as etapas de integração podem ser compreendidas em Mapeamento de Esquemas, Resolução de Entidades e Fusão de Dados, descritas a seguir.

A fase de Mapeamento de Esquemas é responsável por encontrar correspondências entre os registros semanticamente correspondentes dos esquemas das fontes participantes do processo de integração. Com base nas correspondências identificadas, essa fase gera mapeamentos entre os registros das fontes (DONG e SRIVASTAVA, 2015). Na Resolução de Entidades, o objetivo é decidir quais registros de fontes diferentes se referem à mesma entidade (DONG e SRIVASTAVA, 2015). A fusão de dados é a etapa onde se resolvem os conflitos de dados (ALENCAR et al. 2018). Esses são resolvidos com base em estratégias de decisão que escolhem, por exemplo, um valor preferido entre os valores existentes para uma determinada propriedade de registro.

Para facilitar a exposição, a discussão a seguir utilizará um exemplo de integração de informações sobre pacientes de dois hospitais: A e B, utilizando o modelo de dados relacional para as fontes. A Figura 5 mostra o mapeamento dos esquemas locais utilizados. As relações (tabelas) são representadas em itálico e seus atributos aparecem aninhados a elas.

Figura 5: Esquemas locais para o exemplo em uso.



Fonte: As autoras.

Conforme a Figura 5, no esquema do Hospital A, existem três relações (tabelas): Paciente (*r1*), cujos atributos são o nome do paciente (*nome_paciente*), o número do prontuário (*numero_prontuario)* e a data de entrada (*data_entrada*) do paciente no hospital. Observações (*r2*), descrita pelos atributos *descricao* (alguma observação sobre o paciente), *paciente* (*chave estrangeira que possui ligação com a relação Paciente*) e *medico* (chave estrangeira que possui ligação com a relação Médico). Médico (r3), composto pelos atributos: *nome* (nome do médico), *CRM* (número de registro do médico no Conselho de Medicina) e *especialidade* (especialidade do médico). Já no esquema do Hospital B existe apenas uma relação: Paciente (r4), cujos atributos são: nome (nome do paciente), data_nascimento (data de nascimento do paciente), nProntuario (número do prontuario do paciente), data_registro (data de entrada do paciente no hospital), medico (nome do médico) e observação (chave estrangeira que possui ligação com a relação Paciente. As correspondências entre os esquemas são representadas por os identificadores *c1* a *c5.* As Tabelas 1 e 2 a seguir reúnem os dados dos esquemas das relações do paciente para os hospitais A e B respectivamente.

Tabela 1 - Registro do paciente no Hospital A

nome_paciente	José Antonio da Silva
numero_prontuário	000168
data_entrada	11/05/2017

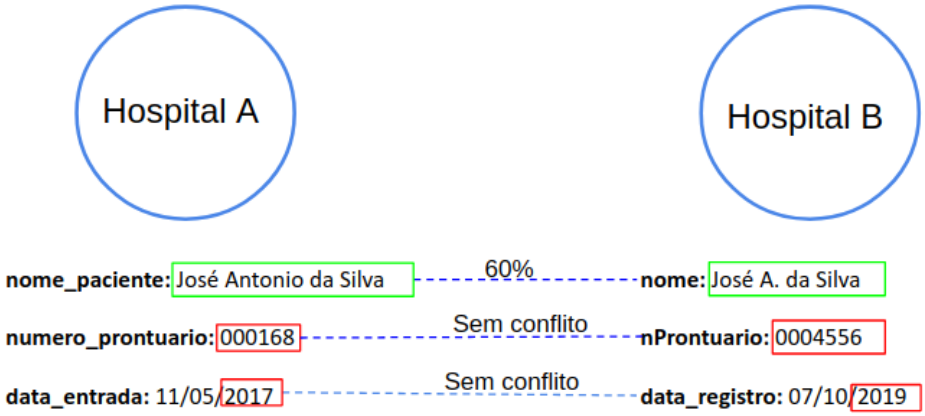
Tabela 2 - Registro do paciente no Hospital B

nome	José A. da Silva
data_nascimento	janeiro-1985
nProntuario	0004556
data_registro	07/10/2019
medico	Fanciley Melsi
observacao	null

Analisando os dados das Tabelas 1 e 2, considerando o mapeamento de esquemas conforme ilustrado através da Figura 5, alguma técnica será aplicada para

identificar que os registros se referem ao mesmo paciente. A Figura 6 mostra um exemplo hipotético de Resolução de entidades, utilizando apenas os dados do paciente.

Figura 6 - Comparação entre registros de um mesmo paciente.



Fonte: As autoras.

De acordo com a Figura 6, são considerados apenas os atributos que mostraram correspondência no processo de Mapeamento de Esquema. Uma técnica utilizada para resolução de entidades é a obtenção de correspondências entre registros. No exemplo, apenas o atributo 'nome_paciente' apresentou correspondências entre os valores dos esquemas dos hospitais A e B. Já os valores dos atributos 'numero_prontuario' e 'data_entrada' (conforme esquema do Hospital A) não apresentaram correspondências significativas entre si, portanto foram considerados como não correspondentes.

Exemplos de técnicas que vêm sendo utilizadas para a Resolução de Entidades são apresentadas a seguir (DONG e SRIVASTAVA, 2015):

- *Blocking*: A ideia básica é aplicar uma *função de bloqueio* (STEORTS, 2014) nos valores de um ou mais atributos para particionar os registros de entrada em vários pequenos blocos e restringir o par correspondente subsequente aos registros no mesmo bloco.

Pairwise Matching: esta etapa compara um par de registros e faz uma decisão local sobre se eles se referem ou não à mesma entidade. Uma variedade de técnicas tem sido proposta para esta etapa, como, por exemplo:

Rule-based: Métodos baseados nessa técnica utilizam os dados dos esquemas das fontes para descobrir informações sobre o conjunto de dados, utilizando para isso métodos de mineração de dados (LI e GAO, 2015).

Classification-based: Essa abordagem utiliza Aprendizado de Máquina Supervisionado (AMS) para criar um classificador para decidir se um par de registros possui equivalência ou não. A vantagem de utilizar abordagens baseadas em AMS é que não exige conhecimento significativo sobre o domínio dos dados. A desvantagem de utilizar essa abordagem é que qualquer tarefa em AMS exige um número grande de exemplos para treinamento do modelo.

Distance-based: Abordagens baseadas em distância aplicam métricas de distância para identificar dissimilaridade dos valores dos atributos correspondentes (por exemplo, usando a distância de Levenstein para calcular dissimilaridade de strings e distância

euclidiana para calcular a dissimilaridade de atributos numéricos) (ELMAGARMID et al., 2007) e utiliza a soma ponderada como a distância em nível de registro.
Clustering: Essa abordagem utiliza o método de clusterização (AM não supervisionado) para criar grupos com os registros que contém algum nível de correspondência ou similaridade. Diferentemente de outras tarefas de Mineração de Dados como a Classificação, na Clusterização os dados de entrada não são rotulados, por isso a Clusterização é uma tarefa de aprendizado não supervisionada. Os registros de um grupo possuem correspondências e similaridades entre si, mas são disjuntos dos registros de outro grupo.

A Fusão de Dados está relacionada ao processo de representar o mesmo objeto, oriundo das diversas fontes de dados envolvidas no processo de integração, em uma representação única, consistente e limpa (BENEVENTANO e BERGAMASCHI, 2019). Algumas técnicas são citadas na literatura no que diz respeito ao processo de Fusão de Dados. De acordo com Dong e Rekatsinas (2018), nos processos iniciais da Fusão de Dados foram utilizados métodos baseados em regras como média e votação e métodos de Mineração de Dados. Atualmente o grande conjunto de trabalhos em Fusão de Dados utiliza principalmente aprendizado não supervisionado para estabelecer correspondências entre os dados e relacioná-los com as fontes (DONG e REKATSINAS, 2018).

3. Trabalhos relacionados em Resolução de Entidades

No que se refere a estudos sobre Resolução de Entidades no domínio de acidentes de trânsito, o trabalho de Kamaluddin et al., (2018) apresenta uma abordagem no contexto de dados de acidentes viários na Malásia provenientes de dados da polícia e dados de hospitais. Sua proposta era realizar a integração de registros de vítimas de acidentes de trânsito registrados pela polícia com as informações das vítimas de acidentes de trânsito registradas nos hospitais. Como diferencial do estudo, identificadores pessoais únicos foram usados para vincular dados de acidentes entre as bases de dados da polícia e do hospital, permitindo a eliminação de falsos positivos. A correspondência de dados foi realizada usando um software que realiza integração semiautomática dos dados apenas na etapa de mapeamento de esquemas, o Microsoft[6] SQL (MSSQL), utilizando a estratégia de pareamento de registros. Nessa abordagem, o software também foi utilizado para facilitar a estimativa da taxa de correspondência com base em atributos autênticos e/ou comuns entre os registros.

Por exemplo, para obter o número de ligações foram utilizadas abordagens determinísticas e probabilísticas. A abordagem determinística foi baseada em correspondências exatas, utilizando como atributos: o número de identificação pessoal e/ou nome e/ou sobrenome das vítimas. A abordagem determinística foi baseada em correspondências exatas do número de identificação pessoal e/ou nome e/ou sobrenome. A abordagem probabilística utilizou uma combinação de asteriscos e/ou caracteres curinga para substituir os caracteres do número de identificação pessoal, nome e/ou sobrenome para garantir a inclusão de grafias e números alternativos (KAMALUDDIN et al., 2018). Como resultados os autores apresentam a conclusão de que os dados da polícia devem ser utilizados com cautela pois, existem diversas inconsistências que, se não forem tratadas de maneira adequada, podem gerar falsos

6 https://www.microsoft.com/pt-br/sql-server/default.aspx

resultados dependendo do escopo da aplicação que os utilizará. Nenhuma análise foi feita em relação à abordagem em si.

Mandacaru et al., (2017) apresentam um estudo para Resolução de Entidades de acidentes de trânsito nas macrorregiões do Brasil. O objetivo deste estudo foi identificar as causas das mortes e estimar o percentual de correções em relação à causa subjacente da morte. O estudo utilizou dados do Sistema de Informações Hospitalares (HIS), do Sistema de Informações sobre Mortalidade (MIS) e do Tráfego Rodoviário Policial. O RecLink III foi usado para realizar a Resolução de Entidades, identificando os pares correspondentes para calcular a porcentagem de correção global da causa subjacente da morte, a circunstância que causou o ferimento no trânsito e a gravidade do ferimento das vítimas no banco de dados policial. O RecLink III utiliza técnicas probabilísticas para descoberta dos relacionamentos entre os registros de uma base de dados, sua avaliação das correspondências é feita com base na identificação de pares. O processo de identificação de pares envolve três etapas: a padronização, a blocagem (*blocking*) e o pareamento de registros. Na fase de padronização, os registros foram organizados de maneira a reduzir as divergências entre os dados. Por exemplo, campos de nomes foram convertidos para maiúsculo ou minúsculo, conforme definição dos especialistas. Na fase de blocagem, os dados foram logicamente divididos em blocos mutuamente exclusivos. Registros pertencentes a um bloco possuem maior probabilidade de representarem pares verdadeiros. Por exemplo, campos como 'Endereço' das vítimas podem ser empregados na blocagem pois serão comparados os registros que apresentem o mesmo endereço. O pareamento de registros consiste na atribuição de escores aos pares de registros originados a partir da etapa de blocagem. O escore serve para classificação de pares como: verdadeiros, falsos ou duvidosos. Um escore acima do limite estabelecido, que é configurado no RecLink III, é considerado como verdadeiro. O limite estabelecido não foi informado no artigo de apresentação do trabalho. Nenhuma análise sobre o método utilizado foi realizada, as conclusões sobre o estudo são em relação às suas contribuições para o campo de pesquisa de Integração de Dados.

O trabalho de Zhao e He (2019) propõe um modelo de transferência de aprendizado para Correspondência de Entidades, utilizando modelos pré-treinados a partir de bases de conhecimento em larga escala. Para cada tipo de entidade, o modelo utiliza sinônimos para rotular e então detectar se pertencem ao mesmo conjunto, utilizando Redes Neurais Hierárquicas. A arquitetura do modelo é composta por três componentes principais: 1) detecção do tipo de atributo, 2) detecção do atributo e 3) detecção em nível de registro. A detecção do tipo de atributo foi realizada utilizando modelos de treinamento off-line, ou seja, não foi realizada em tempo real. O segundo componente, a detecção do atributo, foi realizada tomando como entrada dois valores como, por exemplo, "Dave M. Smith" e "David Smith". Como resultado foi produzida uma pontuação indicando a probabilidade de correspondência para as duas entradas. O terceiro componente é a Correspondência de Entidade em nível registro. Durante essa fase, são utilizados os resultados da etapa de detecção de atributos para auxiliar a resolução de registros, utilizando Redes Neurais pré-treinadas. Como resultados, os autores afirmam que os experimentos realizados em várias tarefas reais sugerem que a abordagem pré-treinada é eficaz e supera os métodos Resolução de Entidades existentes.

O trabalho apresentado neste relatório se diferencia dos demais por utilizar

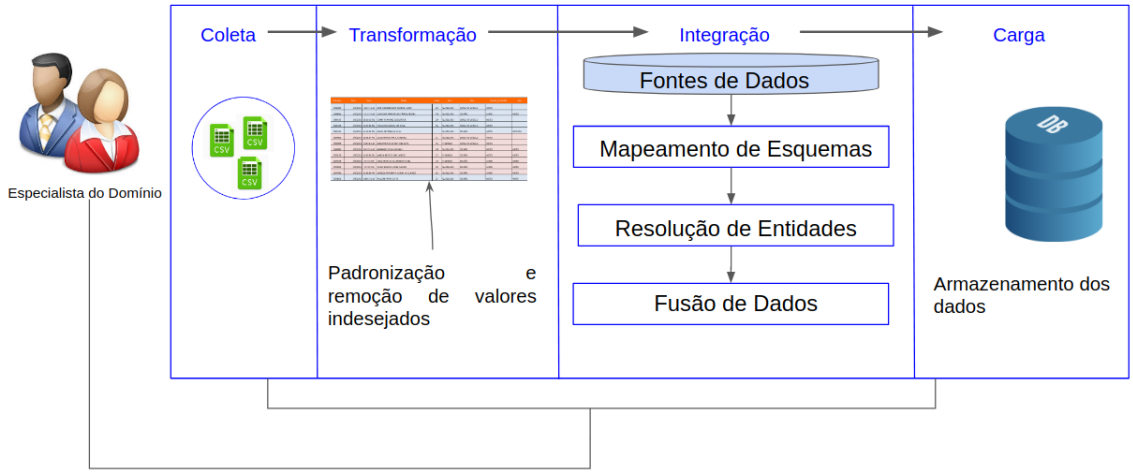
métodos de clusterização para agrupar instâncias que possuem certo grau de correspondências/similaridades. Essa abordagem tem por objetivo ajudar na automatização do trabalho realizado pelos Especialistas de Domínio da SEMOB-JP para encontrar registros duplicados entre os conjuntos de dados sem a necessidade de pré-treinar modelos.

4. Cenário de aplicação e avaliação de algoritmos de clusterização

Este relatório apresenta um estudo acerca de possibilidades de técnicas para resolução de entidades no domínio de acidentes de trânsito. Para isso, foram utilizados dados oficiais de acidentes de trânsito da cidade de João Pessoa, entre o período de 2010 a 2019. Os dados foram fornecidos pela SEMOB, que mensalmente recebe esses dados de outros órgãos como o Serviços de Atendimento Móvel de Urgência (SAMU), hospitais municipais especializados em traumas e a Polícia Civil.

O papel da SEMOB-JP nesse processo é coletar os dados, tratá-los, integrá-los e gerar análises. O resultado dessas análises é utilizado pelos gestores de trânsito da cidade de João Pessoa para tomada de decisões estratégicas. Atualmente, os Especialistas do Domínio de trânsito realizam um processo de ETL, que envolve desde a coleta, integração até o armazenamento, de forma manual e não automatizada. A Figura 7 exibe as etapas do processo de ETL realizadas atualmente pelos Especialistas da SEMOB-JP.

Figura 7: Etapas do processo ETL dos dados de acidentes de trânsito.



Fonte: As autoras.

Na etapa de Coleta, os dados são obtidos a partir de seis fontes diferentes em formato CSV. Na etapa de Transformação, são aplicadas tarefas de limpeza e padronização sobre os dados, por exemplo, remoção de valores nulos e formatação de atributos como data, sexo, etc. Durante a integração propriamente dita, subetapas de mapeamento, resolução de entidades e fusão de dados são executadas. Os Especialistas do Domínio, manualmente, identificam os registros que são correspondentes. Quando dois ou mais registros são identificados como iguais, um dos registros é escolhido para ser armazenado na base de dados. Antes do armazenamento, é feita a integração com os valores do registro escolhido com os dados do(s) registro(s) identificados como correspondentes, caso haja necessidade de adicionar informação para tornar o registro escolhido mais completo. A votação para escolha de um registro é feita com base na confiabilidade da fonte, ou seja, se dois ou mais registros são iguais o que será persistido

é aquele registro presente na fonte tida como mais confiável para o domínio. Por fim, na etapa de Carga, os arquivos transformados são importados para uma base de dados relacional e, a partir de então, são utilizados para geração de relatórios e análises.

Os principais problemas da abordagem descrita são a falta de padronização nos dados, tanto nos atributos quanto nos valores dos registros. Os Especialistas têm de realizar as correspondências manualmente com base no conhecimento do domínio. As Tabelas 3 e 4 apresentam exemplos da falta de padronização dos dados das fontes SAMU e hospital de Traumas respectivamente.

Tabela 3 - Trecho de dados do SAMU em João Pessoa no mês de janeiro de 2015.

Data	Hora	Nome	Idade	Sexo	Tipo	Veiculo_Envolvido	Com	BAIRRO
1/1/2015	3:02:09	DOS SANTOS FILHO	12	MASC	COLISÃO	CARRO	MOTO	VALENTINA
1/1/2015	5:56:15	PESSOA PAIVA	43	FEM	QUEDA DE VEÍCULO	MOTO		X

Tabela 4 - Trecho de dados do hospital de Traumas de João Pessoa no mês de janeiro de 2015.

Data e Hora	Nome	Idade	Sexo	Motivo do atendimento	Detalhe do acidente	Procedência
1/1/15 4:12	DOS SANTOS FILHO	13 anos	MASCULINO	ACIDENTE DE MOTOCICLETA	OUTROS	VALENTINA FIGUEIREDO
1/1/15 6:42	PESSOA DE PAIVA	43 anos	FEMININO	ACIDENTE DE MOTOCICLETA	QUEDA / OUTROS	OITIZEIRO

Conforme demonstrado através das Tabelas 3 e 4, as heterogeneidades encontradas nos dados acontecem em nível de esquema e em nível de instâncias. Dessa forma, a tarefa dos Especialistas é identificar tais correspondências e selecionar quais registros (instâncias) se referem à mesma pessoa.

Resolver problemas de encontrar redundâncias e conflitos nos dados é uma atividade que custa tempo e esforço por parte dos Especialistas. O exemplo descrito neste relatório envolve apenas duas fontes, mas, na realidade da empresa, existem mais quatro fontes envolvidas. A média mensal em cada fonte é de 600 registros, multiplicados por 6 fontes que chega a um total de 3.600 registros para serem analisados pelos Especialistas. Com isso, fica evidente que mesmo após a análise dos Especialistas ainda existem inconsistências nos dados que são replicados para as bases de dados e influenciam negativamente na geração de relatórios para os gestores. A Resolução de Entidades é realizada com o auxílio dos Especialistas.

A análise dos resultados da implementação dos algoritmos de Clusterização foi realizada de acordo com o seguinte aspecto: quantidade de registros corretamente clusterizados de acordo com a avaliação do especialista. A Figura 8 exibe a lógica do algoritmo K-Means.

Figura 8. K-Means - Pseudocodigo

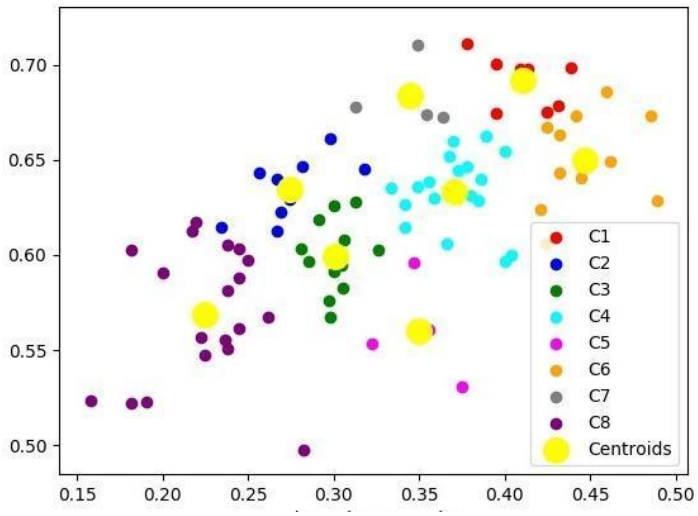
```
Entrada: k (o número de clusters),
         D (um conjunto de elementos)
Saída: um conjunto de clusters k
Método: escolher arbitrariamente k objetos de D como os centros
iniciais de cluster.
Repetir:
     1. reatribuir cada objeto para o cluster no qual o objeto é
mais similar, com base no valor médio de cada objeto no cluster
     2. atualizar o cluster médio, ou seja, calcule o valor
médio dos objetos para cada cluster
Até não haver mais mudanças
```

Fonte: As autoras

O K-Means, na sua forma original, utiliza duas entradas: o número de clusters (k) e o conjunto dos dados (D) a serem agrupados, resultando como saída um conjunto de clusters com os respectivos dados agrupados. Sua execução consiste em calcular o valor médio com base em uma distância definida para cada ponto (dado) do conjunto D e em seguida atribuir esse dado para um cluster cujo valor médio tenha mais semelhança com o valor médio do ponto de dado do conjunto D. Em seguida, o valor médio do cluster é atualizado conforme os valores médios dos objetos pertencentes a cada cluster. Esse processo é repetido até que não haja mais mudanças nos valores médios dos clusters.

Para uma amostra de 312 registros o K-Means agrupou: 42 registros no *cluster1*, 76 registros no *cluster2*, 51 registros no *cluster3*, 3 no *cluster4*, 7 no *cluster5*, 23 no *cluster6*, 37 no *cluster7* e 73 no *cluster8*. Do total de dados agrupados foram encontradas 3 correspondências de registros, e das correspondências encontradas, 2 foram aceitas como corretas pelos Especialista de Domínio da SEMOB, que significa que os registros são os mesmos. A Figura 9 exibe o gráfico de agrupamento dos registros de acidentes, utilizando o algoritmo K-Means.

Figura 9: Gráfico de clusterização gerado pelo K-Means.



Fonte: As autoras.

O gráfico da Figura 9 exibe os *clusters* (C1 a C8) e os pontos de dados gerados pela implementação do K-Means. O algoritmo foi configurado para gerar randomicamente os números de *clusters*, através da extensão K-Means++. O número de iterações do algoritmo foi definido como 100, e os dados de entrada foram o dicionário contendo os registros normalizados na etapa de Transformação. A Figura 10 exibe um

fragmento dos dados clusterizados de maneira mais legível.

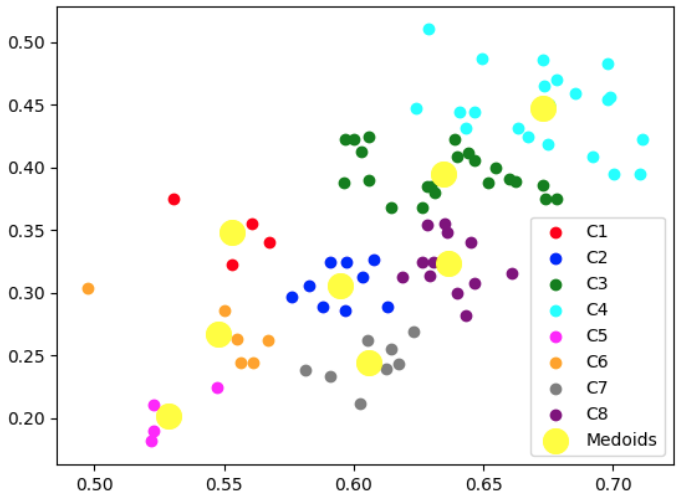
Figura 10: Fragmento dos dados clusterizados pelo K-Means.

Label	Vector	Cluster ID
ADAILTON CORREIA GOMES	1,0	1
ADJAMIR OLIVEIRA DORNELLAS DE CARVALHO	1,50	5
ADRIANO AUGUSTO DA SILVA	1,26	7
ADRIANO FARIAS DE OLIVEIRA	1,21	2
AILTON LOPES DA SILVA FILHO	1,26	7
ALAINE GONÇALVES DA SILVA	0,0	1

Fonte: As autoras.

O gráfico da Figura 11 exibe os *clusters* (C1 a C8) e os pontos de dados gerados pela implementação do K-Medoids, utilizando o mesmo conjunto de dados de entrada. As configurações do K-Medoids foram feitas para serem as mais próximas possíveis do K-Means para não enviesar o resultado da execução do algoritmo.

Figura 11: Gráfico de clusterização gerado pelo K-Medoids.



Fonte: As autoras

Como resultados, para o mesmo número de amostras do K-Means, o K-Medoids agrupou: 11 registros no *cluster1*, 68 registros no *cluster2*, 86 registros no *cluster3*, 83 no *cluster4*, 3 no *cluster5*, 13 no *cluster6*, 17 no *cluster7* e 31 no *cluster8*. Do total de dados agrupados foi encontrada 1 correspondência de valor para a amostra, e a correspondência encontrada, foi aceita como correta e era o valor esperado pelos

Especialistas da SEMOB.

5. Considerações e trabalhos futuros

Este relatório apresentou resultados iniciais de um estudo cujo objetivo foi avaliar a viabilidade de técnicas de clusterização aplicadas a dados de acidentes de trânsito, especialmente, no apoio à tarefa de Resolução de Entidades. Os resultados iniciais obtidos serviram para avaliar o uso de dois algoritmos de clusterização escolhidos. Nesse cenário, ainda não foi possível realizar conclusões significativas que servirão como diretrizes para os trabalhos futuros. Diversas etapas do processo descrito no presente texto têm de ser melhoradas para que a eficiência da automatização do processo de Resolução de Entidades seja aplicada aos processos da SEMOB-JP.

A proposta do estudo inicialmente é investigar e avaliar as técnicas estudadas para então decidir qual empregar na etapa de Resolução de Entidades. Além das técnicas utilizadas nesse estudo, outras serão também estudadas. Após a definição de quais métodos serão utilizados, pretende-se implementar e testar os algoritmos escolhidos utilizando como entrada os dados de acidentes de trânsito de João Pessoa.

Como trabalhos futuros, pretende-se experimentar outros algoritmos de clusterização como o Fuzzy K-Means, C-Means e outros. O objetivo é encontrar um algoritmo que consiga trabalhar com atributos nominais sem a necessidade de normalização dos dados. Após isso, a pesquisa pretende se estender para outras atividades da integração dos dados como a fusão de dados. Pretende-se também implementar uma proposta de execução incremental de clusterização e aumentar o número de amostras para obter melhor acurácia.

6. Referências

ABITEBOUL, S. et al. **Data on the Web**. 1st. Ed. Morgan Kaufmann Publishers, 2000.

ALENCAR, A. et al**. Integrating User Profiles from Academic and Professional Web Data Sources**. In: Anais dos Eventos CLEI-LACLO 2018. São Paulo. 2018.

BANSAL, S., KAGEMANN, S. **Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration**. In. IEEE International Congress on Big Data. DOI:10.1109/BigData.Congress.2014.82, 2014.

BENEVENTANO, D., BERGAMASCHI, S**. Entity Resolution and Data Fusion: an integrated approach**. In: Intervento presentato al convegno SEBD 2019: 27th Italian Symposium on Advanced Database Systems tenutosi a Castiglione Della Pescaia, Italy nel Jun 16, 2019 - Jun 19, 2019.

DOAN, A. et al. **Integration in Support of Collaboration**. 10.1016/B978-0-12-416044-6.00018-1, 2012.

DONG, L. X., SRIVASTAVA, D. **Big Data Integration**. Morgan & Claypool, 2015.

DONG, X., REKATSINAS, T. **Data integration and machine learning: a natural synergy**. Proceedings of the VLDB Endowment. 11. 2094-2097. 10.14778/3229863.3229876, 2018.

ELMAGARMID, A., et al. **Duplicate Record Detection: A Survey.** In IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1-16, Jan, 2007.

KAMALUDDIN, N. et al. **Matching of police and hospital road crash casualty records – a data-linkage study in Malaysia**. International Journal of Injury Control and Safety Promotion. 26. 1-8. 10.1080/17457300.2018.1476385, 2018.

KATSIS, Y., PAPAKONSTANTINOU, Y. **View-Based Data Integration**. In: Liu L., Özsu M.T. (eds) Encyclopedia of Database Systems. Springer, New York, NY, 2018.

LI, L., GAO, H. **Rule-Based Method for Entity Resolution**. Knowledge and Data Engineering, IEEE Transactions on. 27. 250-263. 10.1109/TKDE.2014.2320713, 2015.

LIMA, T. F. et al**. ANÁLISE EPIDEMIOLÓGICA DOS ACIDENTES DE TRÂNSITO NO BRASIL.** Encontro de Extensão, Docência e Iniciação Científica (EEDIC)**,** [S.l.], v. 5, n. 1, mar. 2019. ISSN 2446-6042. Disponível em: <http://publicacoesacademicas.unicatolicaquixada.edu.br/index.php/eedic/article/view/3102>. Acesso em: 14/04/2020.

LÓSCIO, B. F., SALGADO, A. C**. Integração de Dados na Web**, In: Anais da VI Escola Regional de Informática, São Carlos, 2013. Disponível em: Acesso em 13/04/2020.

MANDACARU, P. M. P., et al**. Qualifying information on deaths and serious injuries caused by road traffic in five Brazilian capitals using record linkage**, In: Accident Analysis & Prevention.Volume 106, September 2017, Pages 392-398, 2017.

STEORTS, C. R., et al. **A Comparison of Blocking Methods for Record Linkage**.In: Domingo-Ferrer J. (eds) Privacy in Statistical Databases. PSD 2014. Lecture Notes in Computer Science, vol 8744. Springer, Cham, 2014.

WIEDERHOLD, G. **Mediators in the architecture of future information systems**, IEEE Computer, p.38-49, 1992.

ZHAO, C. e HE, Y. **Auto-EM: End-to-end Fuzzy Entity-Matching using Pre-trained Deep Models and Transfer Learning**. 2413-2424. 10.1145/3308558.3313578, 2019.