UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Marília

William Pires de Castro

Análise de uso de algoritmos de *machine learning* para desambiguação de entidades

Marília
2023

William Pires de Castro

Análise de uso de algoritmos de *machine learning* para desambiguação de entidades

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Infomação como parte das exigências para a obtenção do título de Mestre em Ciência da Informação pela Faculdade de Filosofia e Ciências, Universidade Estadual Paulista (UNESP), Campus de Marília.

Área de Concentração: Informação, Tecnologia e Conhecimento

Orientador (a): Prof. Dr. José Eduardo Santarem Segundo

Marília
2023

C355a

Castro, William Pires de
ANÁLISE DE USO DE ALGORITMOS DE MACHINE LEARNING PARA DESAMBIGUAÇÃO DE ENTIDADES / William Pires de Castro. -- Marília, 2023
65 p.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Filosofia e Ciências, Marília
Orientador: José Eduardo Santarem Segundo

1. Desambiguação da Informação. 2. Ambiguação da informação. 3. Aprendizado de Máquina. 4. Entidades Nomeadas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Filosofia e Ciências, Marília. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

William Pires de Castro

Análise de uso de algoritmos de *machine learning* para desambiguação de entidades

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade Estadual Paulista “Júlio de Mesquita Filho” (Unesp), como requisito parcial para a obtenção do título de Mestre em Ciência da Informação.

Área de concentração: Informação, Tecnologia e Conhecimento
Linha de pesquisa: Informação e Tecnologia

Banca Examinadora

Prof. Dr. José Eduardo Santarem Segundo
UNESP – Câmpus de Marília
Orientador

Prof. Dr. Caio Saraiva Coneglian
UNESP – Câmpus de Marília

Prof. Dr. Fábio Piola Navarro
UNIMAR – Marília

Marília, 15 de março de 2023.

AGRADECIMENTOS

Primeiramente, agradeço a Deus pela companhia em todos os momentos e aos meus pais, Catarina e Wilson pelo apoio e incentivo em todas as escolhas e caminhos que minha jornada tem proporcionado.

A Júlia Vargas, pessoa que vem acompanhando minhas alegrias e percalços da vida nos últimos anos e me motivando em todo processo.

Ao Fábio Navarro, por me mostrar a possibilidade de ingressar em um programa de pós-graduação e me tutorar no início do programa.

Agradeço a minha companheira de pesquisa, e futura Doutora, Ananda de Jesus, que me auxiliou em vários momentos durante a construção da pesquisa.

Ao André Machado, que colaborou com a montagem e construção do texto com correções ortográficas e sugestões de melhorias na escrita e coerência do trabalho.

Por fim ao meu orientador, Prof. Dr. José Eduardo Santarem Segundo, pelas inúmeras orientações ao longo de todo o trabalho, sempre visando a realização da pesquisa da melhor forma possível. Agradeço a oportunidade de realizar o Mestrado acadêmico e por ter me aceito como seu orientado.

RESUMO

O ambiente digital trouxe diversas inovações para a forma com a qual o material científico é consumido. Entretanto, muitas revistas, anais de eventos e afins não se preocupam com a forma de armazenamento dos trabalhos inseridos, permitindo que dados ambíguos sejam cadastrados, como por exemplo as divergências nas abreviaturas de nomes, erros de escrita e atribuições indevidas de trabalhos para autores homônimos, inviabilizando o gerenciamento da base. A área de Desambiguação da Informação estuda formas de se tratar informações ambíguas, contando com técnicas de aprendizado de máquina para desambiguar informação científica. De acordo com o cenário descrito, questiona-se: como a literatura trata a desambiguação de entidades, tais como nomes de autores, utilizando aprendizado de máquina? Esta pesquisa tem como objetivo analisar a abordagem da comunidade científica para a desambiguação de nomes de entidades, buscando compreender a definição dos conceitos da área, identificando as principais formas de execução e lacunas existentes nos métodos de desambiguação avaliados. Quanto à abordagem dos artigos, foram identificadas duas principais divisões: a desambiguação por agrupamento (aprendizado não-supervisionado) e por classificação (aprendizado supervisionado), estendendo-se ao uso de algoritmos para análise dos resultados do processo, visualizando a eficiência do método escolhido. A maior diferença entre os métodos são seus filtros, sendo os mais populares a rede de citações e a rede de co-autoria. Quanto aos desafios científicos, observa-se que a maioria dos textos avaliados sugere a adição de outras formas de desambiguação para ajustar a acurácia, seja por inteligências artificiais bem treinadas ou validação humana dos resultados. Conclui-se que a área de desambiguação de nome de autores tende a processos de agrupamento, mas sem um consenso definido sobre como seguir a partir deste ponto, onde os filtros se tornam a forma principal de distinguir uma pesquisa da outra, podendo levar a novas pesquisas a respeito do assunto.

PALAVRAS-CHAVE: Desambiguação da Informação; Ambiguação da informação; Aprendizado de Máquina; Entidades Nomeadas.

ABSTRACT

Many journals, digital repositories and events have a manual data input of papers, with no previous indexes that carry information from the authors, where ambiguous information can be inserted into the bases, such as divergences in the abbreviations of names, writing errors and allowed attributions of works for authors with homonymous names, making the database management process unfeasible. The Information Disambiguation area has been studying ways to handle similar scenarios, relying on machine learning techniques to disambiguate scientific information. According to the scenario described, the question is: how does the literature treat the disambiguation of entities, such as author names, using machine learning? This research aims to analyze how the Information Science community approaches the disambiguation/ambiguation of entity names, seeking to understand the definition of concepts around the area, identifying the main forms of implementation and gaps in the evaluated disambiguation methods. As for the approach of the articles, two main subdivisions were identified, being disambiguation by grouping (non-supervised learning) and by classification (supervised learning), extending to the use of algorithms for analysis of the process results, visualizing the efficiency of the chosen method. The most important difference between the methods is their filters, the most popular being the citation network and the co-authorship network. Regarding the challenges, it is observed that most of the texts evaluated suggest the addition of other forms of disambiguation to adjust the accuracy, whether by well-trained artificial intelligences, or human validation of the results. It is concluded that the author's name disambiguation area tends to clustering processes, but without a defined consensus on how to proceed from this point, where filters become the main way of distinguishing one search from another, may lead to many new researches on the subject.

KEYWORDS: Information Disambiguation; Information Ambiguation; Machine Learning, Named Entity.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ADC	- Ambiente Digital Científico
AM	- Aprendizado de Máquina
DNA	- Desambiguação de Nomes de Autores
LISTA	- Library, Information Science and Technology Abstracts
ML	- Machine Learning
MSL	- Mapeamento Sistemático da Literatura
OCS	- Open Conference Systems
OJS	- Open Journal Systems
WoS	- Web of Science

SUMÁRIO

1 INTRODUÇÃO

A área de tecnologia vem se desenvolvendo ao longo dos anos e proporcionando às pessoas várias oportunidades de produção e armazenamento de informações nos mais diversos ambientes. Com isso, a forma de se consultar material científico também tem evoluído constantemente, trazendo consigo avanços que permitiram uma constante migração da forma de armazenamento, consulta e análise de pesquisas para um contexto digital, com o uso de eventos, repositórios acadêmicos, revistas e outras formas de divulgação de conhecimento científico, que neste trabalho serão identificados como “ambientes digitais científicos” (ADC).

Com essa migração, pode-se alcançar maior agilidade nos fluxos de pesquisa, já que a amálgama entre massa de artigos disponibilizados e ferramentas (tais como mecanismos de busca, apurados em ADCs) corroborou com melhorias consideráveis para pesquisadores de diversas áreas. Como exemplo dessas melhorias, pode-se mencionar a maior agilidade no processo de pesquisa.

Como muitos ADC trabalham como divulgadores científicos, seus papéis são bem definidos em divulgar artigos em suas publicações mensais, anais de eventos ou afins, nos quais os trabalhos são cadastrados, avaliados e publicados. Desta maneira, o fluxo de compartilhamento acaba não incluindo uma preocupação com formas de armazenamento de informações com o objetivo de se extrair métricas a respeito do conteúdo.

O montante informacional armazenado nestes locais pode auxiliar na construção de índices sobre os pesquisadores e instituições, para a posterior criação de históricos de publicação. Esses históricos têm a capacidade de se comportar como um auxílio para a gestão do ADC, ou até como um catalogador de produções científicas no cenário nacional.

Os índices citados são construídos a partir das informações cadastradas pelos próprios autores no processo de inscrição de um trabalho para a revista ou evento, com o preenchimento de dados como nome do autor, coautores, nome do trabalho, resumo e afins. Esses são os metadados que identificam o texto dentro do cenário do ADC, podendo ser alterados em processos de revisão pelos moderadores do ambiente.

Alves (2010, p. 47) define os metadados como

> elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades; são ainda dados que descrevem outros dados em um sistema de informação, com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação.

Uma vez que os ADCs não imprimem importância no processo de armazenamento, principalmente após a publicação dos trabalhos, problemas recorrentes envolvendo incongruências podem ser adicionados à base. No processo de adição desse conteúdo, informações errôneas podem ser inseridas, como um nome de autor e instituição não indexável, possibilitando a existência de dados duplicados e uma atribuição de documentos indevida, gerando resultados ambíguos, que apresentam mais de um contexto ou significado. Isso atrapalha o processo de extração dos indicadores, gerando dados de autoridade incorretos.

Para Jesus (2021), os dados de autoridade podem ser constituídos por

> dados bibliográficos criados antecipadamente para a padronização das informações de autor e de assunto, esses dados são mantidos em um catálogo de autoridade para serem aplicados na construção das entradas (e posteriormente dos pontos de acesso) dos catálogos bibliográficos, garantindo assim a uniformidade dos mesmos.

No cenário nacional brasileiro, parte dos ADCs não aplica nenhuma regra de índice ou de criação de dados de autoridade na entrada dos metadados, tornando a ambiguação das bases de dados um problema constante. Para que esses metadados possam ser confiáveis, os dados precisam passar por um processo constante de curadoria, resgatando e redefinindo os dados de autoridade relevantes para cada documento, com a intenção de não se afetar a recuperação dos mesmos. Para tal, uma informação que seja elegível para análise pode passar por um processo descrito como “desambiguação da informação”, que traz a ideia de identificar o assunto pertinente à entidade textual escolhida, isolando-o de conteúdos que não façam parte de seu contexto, seja separando ou unificando entidades.

Existe uma pluralidade conceitual na literatura acerca dos processos de desambiguação, que também podem ser denominados como processos de unificação da informação, já que o processo é responsável por segregar informação não relevante e garantir que dados ligados continuem unificados. Nesse contexto,

este trabalho usará o termo “desambiguação da informação” como padrão. Para além da pluralidade conceitual, as práticas do processo de desambiguação também são diversas, sendo abordadas em diversas áreas, inclusive a Ciência da Informação. A própria CI trabalha com formas gerais de se desambiguar informação; todavia, mesmo que esta ocupação possa ser realizada de forma manual, a robustez do material armazenado traz consigo a necessidade de horas a fio de trabalho de um profissional para que uma base possa ser desambiguada por completo.

A aplicação desse processo na informação contida em bases de dados científicos em âmbito nacional pode auxiliar no processo de extração de informações gerenciais a respeito dos mesmos. Eliminando duplicidades de nomes de autores e instituições, autorias indevidas de trabalhos a respeito de publicações de uma determinada instituição e outros problemas ligados a ambiguação informacional, pode-se trazer uma maior garantia estatística nos resultados levantados.

Em complemento, uma vez que todo o conteúdo mencionado já se encontra na forma digital, a utilização de algoritmos de aprendizado de máquina se torna uma opção viável, já que a literatura tem utilizado tais técnicas em muitas frentes, inclusive na própria CI. Salienta-se que, mesmo que os registros não estivessem em formato digital, poderiam ser digitalizados para que os algoritmos pudessem ser aplicados.

Com a evolução da forma com a qual inteligências artificiais têm trabalhado, existe a possibilidade de se trabalhar com a fusão dos conceitos da CI com a Computação para explorar as possibilidades de se aplicar uma desambiguação de forma automática, para maximizar os resultados ligados a essa prática, realizando a desambiguação em uma escala maior e em menos tempo.

Ressaltada a importância da extração de índices e produção de indicadores no contexto científico, e em especial a necessidade de uma maior padronização a nível nacional, assim como o papel da desambiguação na garantia da consistência desses indicadores e, ainda, a possibilidade do uso de técnicas de aprendizado de máquina nestas operações, questiona-se: como a desambiguação automática de entidades vem sendo abordada e realizada? Parte-se da hipótese de que a utilização de um processo de desambiguação automática da informação com o auxílio de técnicas de aprendizado de máquina pode atuar na tratativa documental, já que, uma vez que a entrada de dados seja indexada, o conteúdo previamente salvo precisa ser validado em sequência.

1.1 Problema

Um dos problemas que dificultam a extração de indicadores dos ADC é a presença de informações ambíguas provenientes de falhas no preenchimento manual dos metadados. Já no processo de inscrição manual, para que um novo autor seja adicionado, existe a necessidade de registrar todos os dados do mesmo, não havendo buscas externas para confirmação de identidade. Tal fato permite que se gere informação ambígua, seja por abreviações variadas de um mesmo nome, erros de digitação ou por nomes duplicados. A existência de nomes duplicados, os chamados homônimos (que ocorrem quando mais de uma pessoa carrega o mesmo nome), pode agravar a situação já que o conteúdo científico de um autor pode conflitar com o conteúdo de outro. Disso decorre a necessidade de outras referências para diferenciar uma pessoa da outra, já que tais fatores tornam a obtenção de indicadores como nome de autor, coautores e instituições, citações e referências bibliográficas de difícil identificação.

Dessa forma, averiguar quem é o autor mais citado dentro de publicações de um ambiente digital científico, levantar dados relevantes de um autor, suas ligações com outros autores, artigos, redes de coautoria e afins se torna um processo complexo, partindo da junção de várias fontes de dados, que muitas vezes não têm conexão nativa, com um trabalho manual extenso. Como agravante, a detecção de ambiguações em uma base aumenta a complexidade do processo. Sem um índice, a base informacional a respeito de dado autor se encontra dispersa e possivelmente distribuída erroneamente entre seus homônimos.

Os procedimentos de análise bibliométrica são fortemente afetados pela falta de tratativas apropriadas de desambiguação e unificação da informação nas bases de dados acadêmicos, seja na manutenção da base documental ou no processo de cadastro de novos elementos.

1.2 Objetivos

Esta pesquisa tem como objetivo principal analisar o uso de algoritmos de aprendizado de máquina na realização de desambiguação automática da informação em ambientes digitais científicos focados em autores e instituições.

Listam-se, como objetivos específicos:

- Discutir os conceitos e os processos da desambiguação da informação e aprendizado de máquina;
- Identificar e analisar processo de desambiguação da informação em cenários semelhantes, como repositórios digitais, acervos institucionais e revistas, e como o aprendizado de máquina tem sido usado nesses casos;
- Identificar e discutir modelos de desambiguação automática de nome de autores;
- Propor um modelo de desambiguação da informação.

1.3 Justificativa

A proposta desta pesquisa é motivada por uma análise das formas como as informações de produção científica são cadastradas e armazenadas em ambientes digitais científicos, como o ENANCIB (Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação), que não apresenta formas de recuperação de informações bibliográficas, como por exemplo o nome do autor. Um autor pode publicar adotando o formato “Sobrenome, Primeiro Nome”, ou “Sobrenome, Primeiro Nome, Segundo Nome”, abreviar o “Primeiro Nome” e manter o restante, abreviar o “Sobrenome” mantendo o primeiro, utilizar outros formatos aceitos por revistas e repositórios acadêmicos, ou simplesmente ter problemas com o tamanho de caracteres aceitos para o “nome do autor e coautor”. Enfatiza-se também que toda a linha de produção de um autor pode ser atrelada a outro, devido a homônimos em relatórios ou em processos de recuperação corriqueiros.

As ferramentas de gestão/armazenamento mais utilizadas para eventos, revistas e repositórios, são respectivamente OCS, OJS e DSpace. O manual da Public Knowledge Project define o OCS (Open Conference Systems) como uma solução para gerenciamento e publicação de conferências escolares online, permitindo um gerenciamento altamente flexível. Já o site Public Knowledge Project define que o OJS (Open Journal Systems) é um *software* online para gerenciamento e publicação de revistas (*journals)*, desenvolvido para melhorar o acesso à pesquisa, tornando-se a maior plataforma aberta do ramo. O Portal do Bibliotecário (2018)

define o DSpace como um software aberto com a finalidade de preservar conteúdo digital, transferindo para a instituição que o utiliza custos e responsabilidade de operação, arquivamento e publicação de seu conteúdo científico, para a criação de repositórios acadêmicos próprios. Ambas as ferramentas não atuam com tratativas para desambiguação da informação a respeito dos trabalhos/autores em seu processo de cadastro e em documentos já armazenados.

Mesmo que os processos sejam implementados para evitar problemas de ambiguação durante a inserção de dados, o conteúdo previamente adicionado ainda precisará passar por rotinas de desambiguação. Dessa maneira, esta pesquisa se justifica por traçar um caminho para desenhar a forma com a qual o fluxo de desambiguação automática da informação funciona, em conjunto com a elaboração de um referencial bibliográfico, referente aos resultados do Mapeamento Sistemático da Literatura (MSL), convergindo com melhores caminhos para viabilizar a implementação de um desambiguador de entidades para ambientes digitais científicos.

O trabalho também tem impacto social, trazendo benefícios a pesquisadores da área, ao compilar a forma com as quais pesquisadores têm lidado com o tema e impactando de forma positiva na captação de recursos informacionais e geração de novos estudos ligados ao tema. Outra vertente seria o impacto que a extração de indicadores tem na comunidade acadêmica nacional, para a qual recursos podem ser disponibilizados de forma errônea derivados de levantamentos divergentes quanto aos totalizadores de contribuições científicas de autores e instituições.

1.4 Trabalhos Correlatos

Nesta subseção, serão apresentados alguns trabalhos focados em análises comparativas de formas de desambiguação da informação.

Para a construção dos trabalhos correlatos, buscas em bases de dados nacionais e internacionais como BRAPCI, LISTA, *Google Scholar* e livros foram realizadas, com a finalidade de se encontrar trabalhos que analisam como o aprendizado de máquina tem sido utilizado para desambiguar informação em ambientes digitais científicos.

A ideia proposta consiste em trazer uma visão geral a respeito do uso de diversas ferramentas de aprendizado, mostrando como são empregadas e em quais pontos do processo de desambiguação elas são utilizadas.

Um primeiro trabalho selecionado é o de Monz e Weerkamp (2009), intitulado *A Comparison of Retrieval-Based Hierarchical Clustering Approaches to Person Name Disambiguation*. Os autores se basearam em uma aplicação de desambiguação de nomes baseada em agrupamentos, utilizando diversas formas de ponderação de termos para avaliar seu desempenho em relação ao *Web People Search* (WePS). Para tal, os autores distribuíram WePS diferentes para montagem de algoritmos para desambiguar, aplicando a mesma forma de avaliação para todos. Como resultado, conseguiram identificar que algoritmos de agrupamento hierárquico aglomerativo são os mais consistentes. Posteriormente, concluíram que pequenas mudanças na função de similaridade podem alterar significativamente o resultado, aumentando substancialmente seu desempenho.

Já o trabalho *Term disambiguation techniques based on target document collection for cross-language information retrieval: An empirical comparison of performance between techniques*, de Kishida (2007), tem a premissa de examinar métodos de resolução de ambiguidades em traduções baseadas em coleções de documentos. O trabalho examina métodos para resolver a ambiguidade em traduções, baseando-se apenas em coleções de documentos selecionados, discutindo dois tipos de técnicas de desambiguação definidas pelo autor: 1) desambiguação por Coocorrência estatística nas coleções; e 2) desambiguação baseada em *feedbacks* de pseudo-relevância. Para examinar o desempenho da tradução usando técnicas de desambiguação focadas em coleções de documentos, os autores utilizaram a língua inglesa como intermediária para traduzir termos do alemão para o italiano, anotando termos de coocorrência, realizando contagem e salvamento em um índice externo, para encontrar informações de ocorrência de forma mais breve. Contaram frases que apareciam juntas para definir o total de coocorrências e, por fim, utilizaram o coeficiente do Cosseno, o coeficiente de *Dice* e o coeficiente *Overleap* para definir medidas de proximidade. Concluíram não haver grande diferença na efetividade da aplicação de técnicas de coocorrência ou de *feedbacks,* o método dominante para desambiguação baseada em termos de coocorrência

Este trabalho traz uma abordagem de buscar ambiguidades no geral, trazendo consigo a ideia central de separação por conjuntos, separando os documentos em coleções. Os trabalhos apontados apresentam temas diferentes, mas com soluções com base semelhante, realizando um comparativo entre métodos de agrupamento para desambiguar algum tipo de informação, sejam nomes de autores ou traduções.

As formas com as quais os documentos são analisados, montados, distribuídos e pré-processados são, em suma, formas de prepará-los para algum tipo de classificação. A grande diferença entre a comparação entre métodos de desambiguação apresentados e este trabalho, é que neste trabalho não há o foco em um método específico, criando a possibilidade de se investigar métodos diversos com funcionamentos completamente diferentes, trazendo uma visão geral, abrangente e recente sobre como pesquisadores trabalham com a desambiguação da informação de forma automática.

Apesar de terem sido selecionados como trabalhos correlatos para esta pesquisa, é importante ressaltar que eles não abordam diretamente o tema da desambiguação da informação em geral, mas sim comparativos de algoritmos utilizados em diferentes áreas, como a desambiguação de nomes e traduções. A escolha desses trabalhos teve o propósito de demonstrar como a área de desambiguação tem sido estudada e comparada, mostrando diferentes técnicas e métodos empregados. Contudo, a ausência de trabalhos que se dediquem a trazer uma visão geral da desambiguação mostra a importância da pesquisa nessa área e a necessidade de investigação de diferentes abordagens e algoritmos para solucionar o problema de ambiguidade na informação.

1.5 Estrutura do Trabalho

Este trabalho se divide em oito seções. Na sessão inicial, apresenta-se a introdução, o problema, os objetivos, a justificativa, os trabalhos correlatos e a estrutura do projeto. Em sequência, a Seção 2 detalha os procedimentos metodológicos e o Mapeamento Sistemático da Literatura (MSL) utilizado para desenvolvimento do trabalho. As Seções 3 e 4 se encarregam de desenvolver conceitos relevantes para o trabalho, mostrando, respectivamente, ambiguação e

desambiguação da informação e aprendizado de máquina. Na Seção 5, trabalha-se com o MSL, analisando como a literatura trata a desambiguação de entidades em cenários semelhantes, com enfoque em nomes de autores. A Seção 6 apresenta resultados relacionados à análise feita na Seção 5, mostrando um modelo de desambiguação de entidades científicas. A Seção 7 contempla as considerações finais em conjunto com possibilidades de evolução para projetos futuros, e por fim lista as referências bibliográficas utilizadas para a construção desta dissertação.

2 PROCEDIMENTOS METODOLÓGICOS

Caracterizando-se como uma pesquisa exploratória, o trabalho adotou pesquisa bibliográfica para compor o referencial teórico sobre os assuntos necessários para o entendimento do conteúdo na totalidade. Configura-se também como uma pesquisa descritiva, extraindo conceitos para a identificação dos objetivos da pesquisa.

Como método para o desenvolvimento da pesquisa, optou-se por um Mapeamento Sistemático da Literatura. O mapeamento se caracteriza como uma pesquisa exploratória, descritiva e quali-quantitativa, tendo sido adotado como forma de coleta de dados. O MSL consiste em um mapeamento focado em encontrar características em um escopo fechado de documentos, obtendo-se um aprofundamento maior a respeito do conteúdo tratado pelos trabalhos selecionados, permitindo a extração de conteúdo relevante para a área de pesquisa nichada.

Kitchenham et al. (2010) definem que os MSL têm uma abordagem ampla, objetivando classificar os estudos de uma temática específica com categorias bem definidas, permitindo uma observação estruturada para a temática em questão, buscando assuntos comuns e assuntos pouco explorados. Dessa forma, para a realização do mapeamento proposto, seguiram-se as seguintes etapas:

1. Planejamento: Elaboração de uma análise exploratória, identificação de trabalhos correlatos e preenchimento do protocolo de busca base de orientação da pesquisa;

2. Execução: Busca nas bases de dados e seleção dos documentos;

3. Sumarização: agrupamento dos documentos por semelhança, criação de categorias para classificação dos resultados, sistematização das informações de interesse em imagens e quadros-resumo.

A análise exploratória foi realizada com o objetivo de promover uma maior familiaridade com os conceitos de desambiguação da informação e aprendizado de máquina. Esse levantamento inicial foi realizado em bases de dados nacionais e internacionais como BRAPCI, LISTA e *Google Scholar*. Também foram consultadas fontes primárias de informação clássicas da área, como livros. Essa etapa de análise

exploratória viabilizou o preenchimento do protocolo de busca necessário para orientar a condução do MSL. O protocolo elaborado é apresentado no quadro 1, a seguir.

Quadro 1 - Protocolo de busca do mapeamento sistemático

Protocolo de busca	
Pergunta de pesquisa	Como assegurar a desambiguação de forma automática de nomes de autores?
Objetivos	Identificar formas de se realizar a desambiguação de nomes de autores
Palavras-chave	*name disambiguation, author disambiguation*, desambiguação de nome, desambiguação de autor
Bases de dados consultada	BRAPCI, LISTA
Período abrangido	2012 a 2022
Idiomas	Português, Inglês e Espanhol
Critérios de Inclusão	(I) Aplica técnicas de desambiguação
Critérios de exclusão	(E) - não está em português/inglês/espanhol; (E) - não está no formato estabelecido (artigo); (E) - apenas menciona a temática de interesse; (E) - não é possível ter acesso ao artigo completo
Categorias de análise	(1) Enfoque do artigo (2) definições de desambiguação (3) aplicação da desambiguação (4) lacunas de pesquisa mencionadas
Data da coleta	23/02/2022
Forma de análise dos resultados	Os documentos foram selecionados de acordo com os critérios estabelecidos. A coleta nos documentos aceitos foi realizada seguindo as categorias de análise apresentadas. Foi elaborado um formulário de coleta contendo autor, título, resumo, DOI e publicadora, e os resultados de cada categoria. Os resultados foram categorizados, sendo identificados padrões de temáticas nos documentos, levantadas as lacunas e agrupadas as definições.

Fonte: o Autor (2022)

Após o término da seleção, realizou-se a leitura dos documentos aceitos, o que viabilizou a criação de categorias para agrupar os mesmos de acordo com as suas semelhanças. Para o estabelecimento dessas categorias, foram considerados os objetivos do proposto trabalho e o conteúdo dos artigos. A nomenclatura aplicada

na criação das categorias corresponde à terminologia utilizada pelos autores, e mais centrada no método aplicado por cada trabalho.

3 AMBIGUAÇÃO E DESAMBIGUAÇÃO DA INFORMAÇÃO

Este capítulo apresenta a temática de ambiguação e desambiguação da informação, trazendo definições e uma abordagem geral sobre os assuntos.

3.1 Ambiguação da informação

Para Maculan e Aganette (2017), a definição de ambiguação vem de termos ou expressões que tendem a múltiplas interpretações, podendo ser causada por estruturas gramaticais e lexicais. O Dicionário Online de Português define ambiguação como “Imprecisão; ação de ficar ou de se tornar ambíguo, de possuir múltiplos sentidos ou variados significados: a ambiguação da lei.”. Dessa forma, independentemente de seu contexto, termos que dão margem a interpretações variadas resultam em informações não confiáveis.

Ambiguidades causadas por fatores gramaticais trazem consigo problemas estruturais no próprio processo de elaboração de textos, algo que Maculan e Aganette (2017) descrevem como enunciados com diferentes possibilidades discursivas. Isso pode ser exemplificado na frase “Jim disse a Artanis que havia lutado muito”, na qual a pluralidade de significado se dá pela dificuldade de identificar o sujeito ao qual a frase se refere. As autoras também levantam a possibilidade de haver ambiguidades ligadas ao léxico, relacionando sinonímia, polissemia e homonímia, ou o simples emprego de palavras que possuem significados variados dependendo do seu contexto. O uso de palavras com pronúncias semelhantes (como “fuzil” e “fusível”) já pode ser suficiente para gerar problemas de entendimento. Mais sério é o caso de homônimos, como a palavra “trem” usada em relação a locomotivas, e “trem” usada como gíria comumente utilizada por pessoas da região de Minas Gerais no Brasil.

Para melhor entendimento dos conceitos apresentados, segue a definição de sinonímia, polissemia e homonímia dada pelo Dicionário Online de Português:

- **Sinonímia**: Particularidade das palavras que são sinônimas; relação de sentido entre duas palavras (vocábulos) que possuem significação muito particular ou própria; Análise e/ou teoria a respeito dos sinônimos;

- **Polissemia**: Que apresenta um grande número de significados numa só palavra; cujo significado dependerá do contexto em que a palavra está inserida;
- **Homonímia**: Qualidade do que é homônimo. Semelhança ou igualdade de palavras com diferentes significados.

Brascher APUD (FUCHS, C. Les ambiguïtés du français. Paris: Orphys, 1996. 183p.) agrega aos motivos linguísticos de ambiguação os aspectos morfológicos, sintáticos, predicativos, semânticos e pragmáticos. Segundo o autor, no contexto morfológico, uma palavra pode transitar entre substantivo, adjetivo, verbo, não sendo classificada em uma categoria gramatical. Na ambiguidade pragmática, o enunciado tem um relacionamento mais voltado à situação do interlocutor, voltando-se a elementos externos, como na frase “Sarah vai à guerra”, que se torna aberta quanto à forma de atuação do sujeito Sarah no contexto, trazendo a possibilidade de atuação como combatente e dúvidas sobre o período vigência no local.

Brascher (2002) finaliza afirmando que

> A ambigüidade pode ser ocasionada por diferentes fenômenos lingüísticos situados nos níveis morfológico, lexical, sintático, semântico e pragmático. A solução destes problemas depende do objetivo de um sistema de recuperação da informação e das bases de conhecimento disponíveis neste sistema.

Informações imprecisas podem levar a tomadas de decisão errôneas, que por sua vez podem trazer diversas consequências, principalmente quando do uso de dados ou informação em sistemas digitais para armazenamento, apresentação e recuperação da informação.

Para Brascher (2002):

> Ao encontrar diferentes significados possíveis de serem extraídos de uma frase ou palavra, o sistema de recuperação necessita distinguir um destes significados determinando, segundo o contexto, qual o significado a ser aplicado, obtendo, dessa maneira, maior precisão na resposta dada ao usuário.

Para que a identificação de contexto seja realizada, é preciso realizar um processo conhecido como desambiguação da informação, com o fim de extrair da

informação contextos diversos e agrupar significados semelhantes, tornando, assim, a informação encontrada válida. Esse tema será explorado em mais detalhes na próxima Seção.

3.2 Desambiguação da informação

O dicionário online de português define a desambiguação como: “Eliminação de duplo sentido de uma frase; ação ou processo que retira a ambiguidade de uma estrutura linguística, explicitando os seus reais sentidos”. O processo de desambiguação de entidades textuais é uma área com desenvolvimento crescente em diversos campos do conhecimento. Entretanto, identificar um autor dentro de uma seleção de entidades é um desafio à parte, já que muitos dados podem ser considerados no processo do isolamento e agrupamento dessas informações.

Para Digiampietri e Ferreira (2018), o processo de desambiguação de nomes de autores tem como grande desafio o fato de registros acadêmicos não obterem nomes completos de seus escritores, armazenando comumente o primeiro nome e sobrenome ou sobrenome e inicial do primeiro nome. Digiampietri e Ferreira (2018) continuam, e afirmam que o processo de Desambiguação de Nomes de Autores (DNA) tem a função de criar agrupamentos de registros de um autor específico, podendo disponibilizar o uso dos dados para padronização de bases de dados ou integração de diferentes bases de informação científica.

Embora a facilidade do uso da tecnologia tenha aproximado os pesquisadores do conteúdo, problemas derivados dessas melhorias apareceram. Por exemplo, a necessidade de se listar trabalhos atrelados a um autor, tarefa que ganha complexidade devido à possibilidade de haver várias pessoas homônimas.

Levin (2010) levanta que

> O problema de identificar representações ambíguas ou duplicadas é bastante antigo, tendo sido definido pela primeira vez em 1959 por Newcombe em (NEWCOMBE, et al, 1959), e formalizado dez anos mais tarde por Fellegi e Sunter em (FELLEGI, SUNTER, 1969). Desde então, o problema vem sendo objeto de pesquisa de um grande volume de trabalhos e publicações

O autor complementa que muitas denominações para o problema foram dadas: *Record Linkage*, de Chaudhuri et al. (2003); *Merge/Purge*, de Hernandez e

Stolfo (1995); *Duplicate Detection*, de Weis e Naymann (2006); e *Reference Desambiguation*, de Kalashniov e Mehrotra (2006).

O autor sugere a divisão da pesquisa em DNA em duas categorias:

- Técnicas com foco na qualidade do resultado;
- Técnicas com foco na agilidade de recuperação dos resultados.

As técnicas com foco na qualidade se fundamentam em encontrar o maior número de dados duplicados possível. Trata-se, na definição de Weis e Naumann (2004), de medidas sofisticadas para se encontrar similaridade entre objetos que precisam ser elaborados. Os autores exemplificam ser de extrema importância separar os "candidatos" da "descrição de objetos", sendo a descrição quem define se uma entidade selecionada para a desambiguação é elegível a comparações com outras, e quais dados podem ser utilizados para realizar as comparações. Com as descrições e definições apresentadas, a declaração dos "elementos influenciadores e dependentes" de um determinado candidato é feita, para se formalizar os tipos de relacionamentos entre as entidades e, então, gerar um modelo de medida de similaridade utilizada para classificar os candidatos.

Já as técnicas de foco em agilidade se preocupam com os custos do processo de desambiguação, ou seja, o tempo necessário para que os resultados sejam obtidos. Nestes casos, os trabalhos admitem algumas duplicatas. Levin (2010) agrega que um dos processos mais utilizados para este modo deriva da técnica da blocagem, que funciona dividindo os dados em blocos. Cada objeto é comparado com outros objetos dentro de um bloco. Apresentando um problema análogo ao modelo de similaridade, o modo de blocos apresenta muitos blocos de tamanho reduzido, gerando um cenário semelhante ao das medidas de similaridade, no qual blocos muito pequenos necessitam de uma quantidade de comparações menor para serem avaliados, mas produzem um grande número de falsos positivos. Quando a comparação é realizada por blocos maiores, mais comparações são necessárias, reduzindo a eficiência proposta pelo modelo.

4 APRENDIZADO DE MÁQUINA (*Machine Learning*)

Para Batista (2003), o aprendizado de máquina é uma subárea de pesquisa importante da inteligência artificial, pois lida com a capacidade de aprender – sendo esta essencial para gerar um comportamento inteligente. Cerri e Carvalho (2017) trazem uma definição similar, mas completam dizendo que *softwares* construídos com a capacidade de autoaprendizado são na verdade sistemas alimentados por conjuntos de dados que referenciam casos passados. Os autores se referem à interdisciplinaridade da área de Aprendizado de Máquina (AM), já que se trata de "uma área de pesquisa multidisciplinar que engloba inteligência artificial, probabilidade e estatística, teoria da complexidade computacional, teoria da informação, filosofia, psicologia, neurobiologia, entre outros".

Essa área de estudo levanta diversos questionamentos para sua evolução, contestando o próprio conceito de aprendizado, estudado em vários trabalhos temáticos. Classificações, agrupamento de dados e processos de previsão de acontecimentos baseados em dados já resguardados são as principais formas de execução da área de AM, o que pode levar a uma interpretação mais estatística da área.

Batista (2003) alega que várias abordagens de aprendizado podem ser utilizadas por sistemas computacionais, como a aprendizagem por hábito, aprendizado por instrução, dedução, aprendizado por analogia e aprendizado por indução. Sanches (2003) explica cada um deles da seguinte forma:

Aprendizado por hábito:

Definido pela falta de necessidade de interferências na informação fornecida da parte do aprendiz, sendo uma forma de aprendizado por memorização direta, com a assimilação explícita pelo aprendiz de um conceito específico.

Aprendizado por instrução:

O aprendizado é realizado a partir de uma fonte, seja uma pessoa na figura de um tutor ou um material, como artigos ou fontes bibliográficas, onde o conhecimento é assimilado, mas não memorizado em comparação com o material original, como no aprendizado por hábito.

Aprendizado por dedução:

Existe uma transformação entre o conhecimento obtido e outros previamente armazenados, com a realização de cálculos ou deduções em cadeia da informação presente, com a possibilidade de memorização do resultado de alguns casos necessários.

Aprendizado por analogia:

A modificação de conceitos já assimilados é praticada para adaptar regras existentes, evitando a criação de novas regras ou novos conceitos.

Aprendizado por indução:

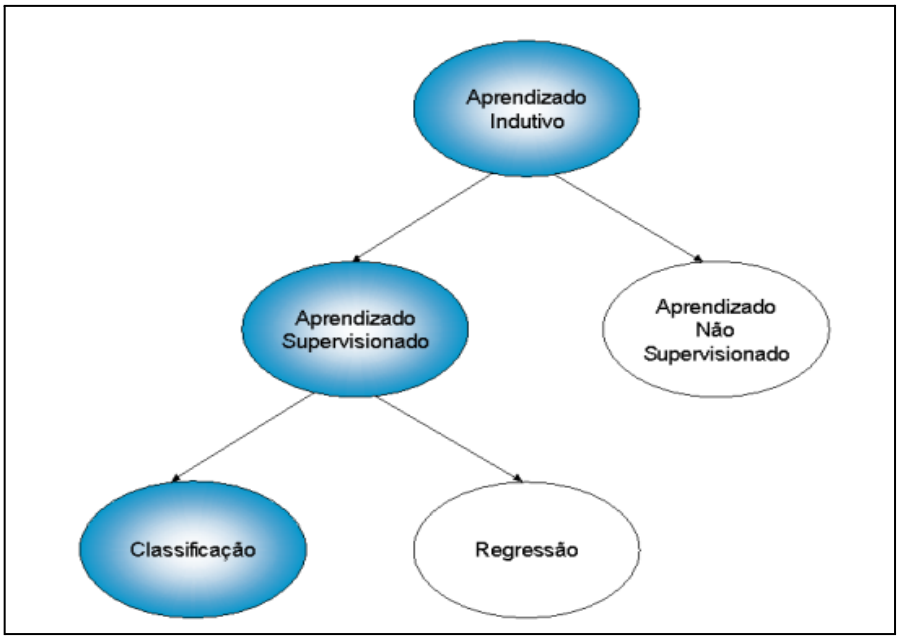
É realizada uma generalização de conceitos a partir de alguns fatos, através da qual padrões entre as coleções são descobertos.

Monard e Baranauskas (2003) completam que o aprendizado por indução trabalha com raciocínio baseado em exemplos, aguardados e fornecidos de forma externa ao sistema de aprendizado em questão. Para os autores, o objetivo desse processo é a construção de um classificador, para identificar corretamente novas entradas ainda não rotuladas. Monard e Baranauskas (2003) trazem que “Na indução, um conceito é aprendido efetuando-se inferência indutiva sobre os exemplos apresentados. Portanto, as hipóteses geradas através da inferência indutiva podem ou não preservar a verdade”.

Já que uma lista de itens é avaliada para cada nova inserção, falsos positivos podem ser apresentados, por demonstrarem semelhanças relevantes. Nesse ponto, vetores de comparação maiores, com quantidade informacional maiores e com fatores de comparação bem definidos ajudam a evitar esses casos.

Monard e Baranauskas (2003) trazem a preocupação no processo de escolha dos atributos envolvidos, já que o método de aprendizagem pode ser sabotado pelos dados utilizados no treinamento. As tarefas preditivas podem ser divididas em tarefas de classificação e tarefas de regressão. Em tarefas de classificação, busca-se atribuir categorias predefinidas a dados.

Figura 1 - Fluxo do aprendizado por Monard e Baranauskas



Fonte: Monard e Baranauskas 2003

Aprendizado Supervisionado

Para Monard e Baranauskas (2003), o algoritmo de aprendizado recebe um aglomerado de dados de exemplo, especificando sua rotulação entre elementos válidos e inválidos, mostrando para o sistema o que considerar para inferir em novos elementos.

Batista (2003) elucida:

> O objetivo do aprendizado supervisionado é induzir um mapeamento geral dos vetores $\vec{x}$ para valores y. Portanto, o sistema de aprendizado deve construir um modelo, y = $f(\vec{x})$, de uma função desconhecida, f, também chamada de função conceito, que permite predizer valores y para exemplos previamente não vistos.

Russel e Norvig (2013) contribuem com duas formas principais de se realizar o aprendizado supervisionado: a regressão, que trabalha com uma forma de aprendizado contínua, e a classificação, que atua com um número finito de resultados.

Batista (2003) complementa trazendo paradigmas propostos para se realizar o aprendizado a partir de conjuntos de exemplos, sendo eles: paradigma simbólico, paradigma estatístico, paradigma *instance-based* e paradigma conexionista.

Paradigma Simbólico:

Com o enfoque de aprender a partir da construção de “representações simbólicas” de um conceito, este paradigma trabalha com análises de exemplos positivos e negativos, que podem ser apresentadas em formas de árvores de decisão, regras de decisão ou redes semânticas. Bastia agrega que as representações simbólicas com maior nível de estudo são as árvores e regras de decisão.

Paradigma Estatístico:

Para Batista (2003):

> Como regra geral, técnicas estatísticas tendem a focar tarefas em que todos os atributos tem valores contínuos ou ordinais. Muitos deles também são paramétricos, assumindo alguma forma de modelo ou distribuição, e então encontrando valores apropriados para os parâmetros do modelo a partir de dados.

O autor contribui comentando a respeito de classificados estatísticos, que podem trabalhar como formas de se encontrar média, variância e covariância de distribuições.

Paradigma *Instance-Based*:

O autor traz como características principais do *Instance Based* o fato de que todos os casos utilizados para treinamento precisam ser memorizados. Isso pode levar a lentidão e dificuldades de manuseio.

Em resumo, esse paradigma atua como um classificador, considera algo já conhecido e assume que os novos casos terão a mesma forma ou classe. Dessa forma, são utilizados casos armazenados para classificar uma nova ocorrência, abrindo precedentes para o uso de um caso específico, mais próximo à nova inserção, ou utilizar vários casos, com diferentes graus de similaridade.

Paradigma Conexionista

O Paradigma Conexionista utiliza redes neurais, baseadas em modelos biológicos do sistema nervoso. O nome conexionista vem da representação interconectada entre todas as unidades.

Aprendizado Não-Supervisionado

O aprendizado não supervisionado vem da premissa de que o usuário não precisa acompanhar nem alimentar o processo com informações iniciais, trazendo uma aproximação maior com a forma na qual o cérebro humano assimila informação a partir das próprias experiências. Sanches (2003) cita que a tarefa do algoritmo de aprendizado não supervisionado é agrupar elementos não rotulados, não tento atributos especificados.

Batista (2013) complementa:

> No aprendizado não supervisionado é fornecido ao sistema de aprendizado um conjunto de exemplos E, no qual cada exemplo consiste somente de vetores $\vec{x}$, não incluindo a informação sobre a classe y. O Objetivo no aprendizado não supervisionado é construir um modelo que procura por regularidades nos exemplos, formando agrupamentos ou clusters de exemplos com características similares.

Dessa forma, a principal diferença entre os algoritmos supervisionados e os não supervisionados é que o primeiro tipo parte de um ponto inicial, onde existem informações rotuladas para o tratamento de novas inserções, já o segundo tipo atua de forma mais livre, criando agrupamentos entre itens que sejam mais semelhantes, consistindo em uma ferramenta considerável para tratar dados não estruturados abundantes.

Russel e Norvig (2013) explicam que os modelos de aprendizado não supervisionado podem operar com uma pretensão de conhecimento a respeito dos parâmetros de resposta, para depois se deduzir a probabilidade de cada elemento pertencer a cada grupo criado pelo modelo. Esse processo se repete até que os elementos se encaixam nos grupos gerados, com atualizações dos grupos a cada iteração.

Aprendizado por Reforço

Para Russel e Norvig (2013), “pode-se considerar que a aprendizagem por reforço abrange toda a IA: um agente é colocado em um ambiente e tem de aprender a se comportar com sucesso nesse ambiente”. Costa (2004) completa que, para ser categorizado como aprendizado por reforço, um agente aprendiz precisa interagir com o ambiente em questão, aprendendo de forma autônoma, sem a ajuda

de nenhum processo de treinamento ou supervisão, a atuar de maneira eficiente sobre um problema.

Esse modelo tem sido empregado em áreas como a robótica e a automação de jogos eletrônicos, com o intuito de trazer proficiência para um agente em um ambiente desconhecido, contando apenas com dados gerados pelo próprio agente e com recompensas sazonais.

4.1 Algoritmos de Aprendizado de Máquina

Este subtópico da Seção apresenta breves definições a respeito dos algoritmos de Aprendizado de Máquina recuperados a partir da análise dos trabalhos aprovados pelo MSL.

Single Linkage Hierarchical:

Para Bonthu (2021), o processo de *Single Linkage Hierachical* se inicia com um tratamento de observação e agrupamentos individuais, para poder mesclar agrupamentos de forma interativa, até que toda a informação seja mesclada em um único agrupamento. Bonthu prossegue, dizendo que o processo de mesclagem é realizado a partir da distância entre os agrupamentos, a qual pode ser calculada a partir de critérios como o *Single Linkage* – o qual, por sua vez, calcula a distância entre dois agrupamentos a partir da distância mínima entre itens de cada um.

F-Measure

Segundo Dembczynski *et al* (2011), o *F-Measure* tem sua aplicação de forma rotineira na extração de métricas de desempenho para fluxos de previsão, como a classificação binária e textos não estruturados e reconhecimento de entidades nomeadas, atuando como uma métrica entre precisão e recuo.

A finalidade do algoritmo é apontar um número único para definir a qualidade de um modelo, trabalhando com conjuntos diversos de dados desproporcionais. Quando o valor da métrica está alto, demonstra relevância. Dessa maneira, quanto maior a quantidade de informação inclusa, melhor a atuação do modelo.

Inverse Document Frequency

Para Kido, Barbon Junior e Moriguchi (2014), o algoritmo de IDF tem sua aplicação voltada ao processo de descobrir palavras importantes em um texto, independentemente da estrutura desse texto (não estruturado ou semi-estruturado). O algoritmo atribui um valor para os termos encontrados e considera sua frequência no texto analisado, bem como em todos os textos disponíveis em uma possível base.

Os autores completam que palavras comuns em grupos pequenos de documentos têm a tendência de ter uma pontuação maior em comparação com pronomes e preposições, que são palavras com grande volume de ocorrência.

Gradient Boosting

Silva (2020) define o algoritmo de *Gradient Boosting* como uma técnica de aprendizado de máquina focada na resolução de problemas de regressão e classificação. O objetivo de seu uso é a construção de um modelo de previsão constituído de modelos de previsão menores e mais fracos, como árvores de decisão.

O autor afirma que, ao gerar uma cadeia de modelos menores, uma cadeia pode remover e minimizar erros dos modelos anteriores, multiplicando a taxa de aprendizagem, valor dado para determinar qual é o impacto que cada cadeia no modelo final. Valores baixos de taxa de aprendizagem indicam pouco impacto no geral.

Naïve Bayes

Gusmão (2022) define o algoritmo de *Naïve Bayes* como um gerador de tabela de probabilidades a partir de um classificador de dados. O algoritmo em questão é oriundo da estatística, mas é muito utilizado para aprendizado de máquina.

A autora segue explicando que o algoritmo permite realizar o aprendizado com múltiplos elementos de forma integrada ou separada, com mais opções de saída para o analisador, podendo ser aplicado a grandes volumes de dados.

Logistic Regression (Regressão Logística)

Para Nick e Campbell (2007), o modelo de regressão logística pode ser definido como “modelos estatísticos que descrevem um relacionamento entre uma variável qualitativa e um valor independente”. Seu uso é voltado ao “estudo do efeito

de variáveis preditoras em resultados categóricos", apresentando resultados binários.

Random Forest

ICSM Júnior (2021) traz a definição do algoritmo *Random Forest* como um algoritmo de aprendizado de máquina voltado a predições, a partir de várias árvores de decisão, que se afunilam para o resultado.

O autor acrescenta que o algoritmo escolhe aleatoriamente duas ou mais variáveis para aplicar cálculos a partir das saídas esperadas, para então definir qual variável será escolhida para ser utilizada no primeiro nó das árvores de decisão, repetindo o processo a cada nó.

Pairwise-F:

A documentação PopGen.jl aponta o algoritmo de *Pairwise-F* como uma evolução do índice de fixação F de *Wright*, baseado em descrever a divergência entre grupos pré-definidos, ou seja, entender o quão misturado estes grupos estão.

Para calcular os pares entre as populações definidas, é preciso realizar um teste de permutação unicaudal com a finalidade de se recuperar os valores P de significância estatística. Para tal, definem-se interações para um valor maior que 0, aplicando um teste de locus (significância ignorada).

5. MAPEAMENTO SISTEMÁTICO DA LITERATURA A RESPEITO DE DESAMBIGUAÇÃO DE NOME DE AUTORES

Nesta seção, conduz-se uma análise de acordo com o conteúdo extraído do MSL, seguindo os procedimentos apresentados na seção de procedimentos metodológicos.

As buscas na BRAPCI retornaram 3 documentos, dos quais apenas um apresentou características que o encaixam no Critério de Inclusão apresentado na Seção 2. Isto posto, é válido levantar a falta de conteúdo nacional relevante da Ciência da Informação sobre a temática de estudo, sendo a desambiguação de nomes de autores ou de entidades no geral. Mesmo utilizando outros termos de busca, como “desambiguação de textos”, “desambiguação de entidades” e afins, a quantidade de documentos levantada pela base em questão não apresentou alteração.

Já as buscas realizadas na plataforma LISTA retornaram 44 trabalhos, sendo que 18 dos documentos recuperados foram aprovados e 26 reprovados.

Ao observar mais profundamente os documentos excluídos, constatou-se que apenas um (1) documento não estava escrito em um dos idiomas selecionados (português/inglês/espanhol); apenas doze (12) artigos mencionaram a temática de interesse; e dois (2) trabalhos não foram encontrados de forma completa. Por fim, treze (13) não falaram a respeito de desambiguação.

Os documentos rejeitados baseados no fator de exclusão “Apenas mencionam a temática de interesse” citam o uso de formas de desambiguação de nomes, mas não necessariamente apresentam uma forma de fato, ou então apresentam formas diversas de desambiguação, mas não sua aplicação.

Abaixo, segue uma breve descrição a respeito dos trabalhos aprovados para levantamento das categorias:

Artigo 1: *A boosted-trees method for name disambiguation*

Propõe um classificador de artigos entre verdadeiro e falso, filtrando por nome e afiliação, para construir um *score* de similaridade. O trabalho também utiliza a *boosted trees classification* para realizar uma classificação, calculando a taxa de erros na técnica. Para tal, foi realizada uma classificação de características em

comum entre os artigos, para que quanto mais características em comum os artigos tivessem, menor fosse a distância entre eles.

Os autores criaram uma base de dados para testes, baseada em um cenário de mundo real, de 4.253 trabalhos de autoria de 100 cientistas norte-americanos sortidos. No processo de criação da base de dados, os registros foram baixados da *Web of Science*, utilizando nome e afiliação.

Para seguir com o processo, uma matriz de similaridade de cosseno par a par foi construída com base no nome do autor. Em seguida, uma decomposição de Eigen foi utilizada para a criação da pontuação de similaridade numérica, permitindo estabelecer a distância entre trabalhos com o mesmo nome de autor.

Artigo 2: *A Fast Method Based on Multiple Clustering for Name Disambiguation in Bibliographic Citations*

Os autores propõem uma estrutura focada em título e local de publicação. O texto, dividido em 3 etapas, trabalha com agrupamentos múltiplos: 1) agrupamento de artigos da rede de coautoria para extração de fragmentos; 2) agrupamento da informação extraída em título; 3) agrupamento das informações resultantes do passo 2 por relações latentes utilizando o algoritmo de *Fast Multiple Clustering*.

O trabalho analisou outras formas de agrupamento, como *Pairwise-F, K-Metric, B-Cubed* e *Cluster-F*, propondo um *framework* para calcular a qualidade dos dados desambiguados, já que as entidades de títulos dos trabalhos e os *clusters* gerados por outros algoritmos podem ser comparados de *cluster* para *cluster*, dispensando a utilização de soluções de “força bruta”. A comparação entre *clusters* utiliza duas tabelas de *hash*, gravando índice de agrupamentos previstos e suas frequências de interseção de agrupamentos com resultados reais.

Em outras palavras, este trabalho traz a ideia de um modelo para cálculo de qualidade das informações desambiguadas a partir da junção de vários algoritmos, comparando agrupamentos realizados por todos.

Artigo 3: *A Graph Combination with Edge Pruning-Based Approach for Author Name Disambiguation*

O trabalho tem a proposta de criar um *dataset* inicial visando aplicar buscas em fontes externas para melhorar a desambiguação focado em nomes homônimos, trabalhando quase que exclusivamente com *Research Gate* e *Google Scholar*, abrangendo possíveis páginas *web* do autor.

Para a construção do *dataset* de consulta, um robô extrator foi construído, consumindo os motores de busca já citados para retirar informações pertinentes para gerar um grafo de similaridade de autor, agrupado com o resultado de uma aplicação de similaridade do cosseno nos *abstracts* dos trabalhos.

O projeto se baseou na *Arnetminer Citation* (AC) e *Web of Science* (WoS) para avaliar os resultados, utilizando algoritmos como o *Pairwise-F* para mensurar a precisão.

Artigo 4: ***A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering***

O trabalho visa propor uma estrutura baseada em camadas, sendo: 1) camada de agrupamento e e-mail; 2) camada de agrupamento por rede de coautoria; 3) camada de agrupamento de artigos.

Adotando um mecanismo dinâmico de agrupamento para minimizar erros e utilizando dados baseados em cenários reais, é possível utilizar metodologias diferentes em cada camada. Assim, pode-se anexar a estrutura em camadas adicionais facilmente, ou conectar a estrutura com outros sistemas externos.

O *framework* permite adicionar e remover camadas usando um mecanismo de conjunto para minimizar erros de agrupamento. Para tal, aplica o algoritmo *single linkage hierarquical* para separar grupos de coautor e títulos para início processo.

Artigo 5: ***A parser for authority control of author names in bibliographic records***

Trabalha com uma ferramenta de código aberto para auxílio de controle de autoridade em catálogos bibliográficos, baseada em semelhanças medidas entre variantes de nome de autores e datas relacionadas ao autor, utilizando um modelo de *unigram frequency vector trie* para acelerar a identificação das variações

Os autores definiram um atributo "grammar", responsável por validar e analisar os dados e dar a eles uma interpretação não-ambígua por um certo período, usando como auxílio na compatibilidade entre os autores dados como ano, século e outras associações referentes ao período de publicação.

A comparação de nomes dos autores é baseada em uma simples comparação entre palavras e, segundo os pesquisadores, emprega uma estrutura de dados compacta para acelerar a pesquisa por variações de nomes.

Artigo 6: ***Author disambiguation using multi-aspect similarity indicators***

Os autores realizaram uma combinação autor/publicação para atribuir um identificador único para cada trabalho analisado, garantindo que cada artigo garanta uma unicidade para o autor do mesmo no início do processo de desambiguação.

Utilizaram também uma seleção de metadados extraídos da *Web of Science* como: título, resumo, rede de citações, palavras-chave, rede de coautoria, endereço, periódicos, diferença entre o tempo das publicações e média de publicações do autor.

Para tal, usaram um cálculo baseado em similaridade dinâmica, onde o método aborda os problemas de registros descartados devido a campos de dados nulos e seu efeito resultante em resultados de *recall* e precisão, com o algoritmo de F-Measure.

Artigo 7: ***Author Name Disambiguation for PubMed***

Os pesquisadores desenvolveram um sistema baseado na estimativa de similaridade e categorização aglomerativa, utilizando um método de *machine learning* para distribuir pontuações. O método considera a compatibilidade entre os nomes dos autores.

Os autores destacam a capacidade de detectar conflitos com variantes de nomes, um esquema de compatibilidade de nomes que facilita a criação e precisão dos dados de treinamento em grande escala e regula o algoritmo de agrupamento *Inverse Document Frequency* de modo a minimizar os agrupamentos falsos positivos.

Dessa forma, os recursos permitem a computação de *scores* de semelhança de pares para estimar a capacidade de um par de artigos pertencentes ao mesmo autor, que impulsiona um algoritmo de agrupamento aglomerativo regulado por compatibilidade de nomes e nível de probabilidade.

Artigo 8: ***Author name disambiguation in scientific collaboration and mobility cases***

Detectando nomes homônimos, o algoritmo desenvolvido pelos autores classifica artigos de um autor, utilizando suas informações de filiação.

Posteriormente, foram classificados artigos com uma rede de coautoria completamente diferente. Isso para que, em um terceiro passo, artigos cujos autores mudaram de filiações pudessem ser classificados corretamente, baseando-se em informações de coautoria.

Para isto, os autores trabalharam em um método baseado em reforço recursivo entre afiliação e coautoria, com múltiplas interações, de modo a, por fim, separar os trabalhos em listas de autoria e coautoria para realizar um cruzamento posterior.

A checagem de resultados foi feita através de um método de amostragem randômica visando selecionar o conteúdo avaliado e, assim, realizar uma verificação manual, comparando-os com os resultados obtidos pelo algoritmo.

Artigo 9: *Author name disambiguation using a graph model with node splitting and merging based on bibliographic information*

Os autores propõem uma abordagem baseada em grafos para o processo de desambiguação, utilizando nomes dos autores e relações de coautoria.

Para isso, criaram um *framework* chamado de *Graph Framework for Author Disambiguation (GFAD),* no qual os problemas de nomes homônimos são resolvidos por meio de uma divisão de um vértice de autores envolvidos em múltiplos ciclos de coautoria. A heterogeneidade dos nomes é garantida pela fusão de múltiplos vértices de autores que tenham nomes similares, caso esses vértices sejam conectados a uns vértices em comum.

A parte de nomes homônimos é resolvida a partir dos círculos sociais de cada um, baseando-se no período e na comunidade de autores ligados a ele. Se um autor estiver ligado a múltiplos círculos sociais, o modelo entende que existe mais de um autor ligado a este nome, separando a informação em vértices diferentes.

A detecção de círculos acontece investigando vértices com múltiplos círculos sociais no grafo a partir da rede de coautoria. No caso de um autor com nomes diferentes, a resolução se dá a partir de uma combinação de vértices relacionadas a este autor.

Por fim, os algoritmos de aprendizado de máquina K-metric, Pairwise-F1 e cluster-F1 foram utilizados para mensurar os resultados.

Artigo 10: *Disambiguation of author entities in ADS using supervised learning and graph theory methods.*

Apresenta um algoritmo para desambiguar autorias no *Astrophysics Data System* (ADS), utilizando uma abordagem semi-supervisionada para treinar um classificador de *tree-based,* com a finalidade de detectar comunidades a partir dos gráficos ponderados.

Os autores combinaram métodos de aprendizado de máquina supervisionado e teoria dos grafos nos registros bibliográficos da *ADS,* aplicando primeiro um bloqueio, agrupando todas as autorias que compartilham o mesmo sobrenome.

A primeira parte acontece após passarem por um processo de pré-processamento, removendo letras minúsculas e caracteres especiais.

Como segunda parte, foi realizado o treinamento do modelo de classificação probabilística, com a finalidade de decidir se dois artigos em um determinado bloco pertencem ao mesmo autor.

Como passo final, é utilizado o classificador probabilístico treinado para criar entidades que vão representar o autor, para utilizá-las como sub gráficos de autoria. O classificador decidia se dois artigos dentro de um bloco pertenciam ao mesmo autor, e posteriormente criava “entidades" baseadas nos autores e grafos de autoria.

Os autores realizaram testes com vários algoritmos de classificação, dando preferência para modelos não lineares baseados em árvore, por terem um melhor desempenho em casos de excesso de dados. Estes testes mostraram que o algoritmo de *Random Forest* obteve o melhor desempenho dentre os testados.

Artigo 11: DISC: Disambiguating homonyms using graph structural clustering

Os autores construíram um gráfico de coautorias utilizando autores e seus respectivos coautores. Posteriormente, foi realizado um agrupamento estrutural de grafos baseado em *gSkeletonClu* para identificar informações relevantes em cada nó do grafo de coautoria.

Nomes homônimos foram resolvidos dividindo os nós categorizados pelos seus vetores de similaridade. Para classificar os resultados, utilizaram algoritmos como Kmetric e PF1.

Artigo 12: Ethnicity-based name partitioning for author name disambiguation using supervised machine learning

O trabalho explora o potencial da etnia do nome, comparando o desempenho de quatro algoritmos de aprendizagem de máquina, treinados e testados com todos os dados ou, especificamente, em grupos de nomes individuais, conseguindo resultados positivos.

Os artigos foram classificados como “*match*” e “*nonmatch*”, agrupando instâncias de nomes em blocos com a mesma inicial e mesmo sobrenome e comparando com outras similaridades como nome de coautores, de modo a produzir uma pontuação de similaridade.

Os algoritmos de *Machine Learning* combinam as pontuações para aprender como “pesar” cada uma, e decidir se fazem parte da mesma instância de um autor ou não. Os autores realizaram testes com 4 modelos de ML (*Gradient Boosting, Naive Bayes, Logistic Regression e Random Forest*), classificando nomes por etnia, considerando o cálculo de precisão e recuo.

Artigo 13: *Evaluating author name disambiguation for digital libraries: a case of DBLP*

Este estudo trabalha com uma abordagem de triangulação de nomes do autor para bibliotecas digitais, testado em três tipos de base de dados, contendo de 5000 a 6000 nomes diferentes.

A primeira base aplicada foi a *Automatically labeled data*, que consiste em separar pares de nomes com base em autocitação, podendo ser utilizada para mensurar o retorno e a quantidade de pares compatíveis.

Outra possibilidade é a aplicação por instâncias do ORCID, através da qual são gerados dois tipos de base de dados com nomes ambíguos: 1) nomes conectados aos ids do ORCID (compatíveis com, no mínimo, outro nome) pertencem a blocos representando dois ou mais autores distintos. 2) nomes conectados aos ids do ORCID e que não se igualam a, no mínimo, um nome estão em blocos de representação distintos.

Por fim, há o *Manually labeled data*, método no qual cada pontuação e a sua média são calculadas por bloco, apresentando seu desvio padrão. Os dados obtidos nesse processo são usados para gerar métricas de precisão trabalhadas com o algoritmo de *Pairwise-F*.

Artigo 14: *Exploiting similarities across multiple dimensions for author name disambiguation*

O artigo propõe um método que usa duas relações independentes entre coautoria e metadados dos documentos, visando gerar uma representação latente de artigos, com a capacidade de generalizar dados sobre vários conjuntos de informação.

Os autores utilizaram o *hierarchical agglomerative clustering* como um método de aprendizado de máquina não supervisionado, combinando aspectos relacionais e não relacionais das coautorias e metadados, gerando representações por dimensão, e gerando um vetor de fusão de informações. Isso permitiu uma melhora na precisão e o uso de técnicas de *representation learning* para mapear registros que incorporam espaços passíveis de serem usados na determinação os agrupamentos.

Dessa forma, os autores propuseram um modelo/método que usa duas relações independentes: 1) coautoria e metadados para gerar uma representação latente do documento, generalizando vários *datasets*; 2) criaram um grafo de similaridade de meta conteúdo, removendo *stopwords*, *lemmatization*, para pré-processar e realizar cálculos de medida de similaridade.

Artigo 15: ***Generating automatically labeled data for author name disambiguation: an iterative clustering method***

O artigo mostra que dados dados podem ser gerados automaticamente pelo uso da tecnologia dos registros da publicação, como o endereço de e-mail, rede de coautoria e referências bibliográficas.

Com o intuito de criar identidades, este trabalho selecionou dados e conectou os resultados em grupos para serem analisados, buscando elementos dos trabalhos na plataforma do ORCID e atribuindo o ID ao nome do autor, caso encontrasse algo satisfatório.

Isso posto, os autores definiram regras de alta precisão para combinar instâncias de nomes, agrupando as instâncias de nomes ambíguos num processo de emparelhamento de pares com base nas regras. Como encerramento, os conjuntos são fundidos por um algoritmo genérico de resolução de entidade. O processo é repetido até que as fusões não sejam mais possíveis.

Artigo 16: ***Identifying author–inventors from Spain: methods and a first insight into results***

Utilizando um procedimento de duas etapas, o artigo se inspira fortemente em desarmes não supervisionados bem estabelecidos, utilizando métodos de ambiguação propostos na literatura, seguindo uma abordagem baseada em rede e exigindo um conjunto restrito de atributos de registro, considerando costumes de escrita de nomes e dicionários para identificar o nome de pessoas e instituições.

Esse objetivo é alcançado estruturando o texto para detectar nomes de pessoas, instituições, endereços, para atrelá-los a outros campos pré-definidos. Os autores trabalharam com reconhecimento de linguagem, divisão de sentenças e fluxos de correção ortográficas para então, classificar os elementos em categorias distintas Também usaram um algoritmo de *string matching* para combinar a fonética, se baseando no estilo de teclado QWERTY para inglês e espanhol.

Para classificar os registros, usaram *bdscan*, baseado em técnicas linguísticas com o *grammar based* para extrair informações. Para classificar dados extraídos, usaram uma classificação de hierarquias "*sub entities*".

Artigo 17: *Incremental Author Name Disambiguation by Exploiting Domain-Specific Heuristics*

Faz a remoção de palavras conectivas, como “a, e, de” e afins, retirando termos relevantes do texto como nome de autores, publicação e título, tratando tudo como um termo a ser analisado.

Os autores trabalharam definindo a similaridade entre conjunto de autores e citações, verificando diferenças entre cada atributo das citações (os atributos em questão sendo coautoria, títulos e local de publicação).

Verificações nos próprios conjuntos de autores também foram realizadas, de modo a separar autores com maior quantidade de publicações. Heurísticas de domínio específico foram utilizadas para identificar a presença de novos autores. Correções de classificações inválidas, usando a similaridade obtida no processo de desambiguação, também foram aplicadas, através do processo de *Fragment Comparison, para* escolher o melhor conjunto.

Artigo 18: *Institution name disambiguation for research assessment*

O trabalho utiliza técnicas existentes para desambiguação, com base na semelhança de palavras ou distância de edição, um algoritmo baseado em regras. O algoritmo é avaliado em grandes conjuntos de dados, com uma boa precisão, mas precisa de melhorias.

Em sua aplicação, os autores seguiram três passos:

1. Construção de uma tabela de autor x instituição;

2. Realização de um reconhecimento dos candidatos a partir de blocos de autores;

3. Base de frequência gerada a partir de uma relação de nome da instituição.

Os autores levantaram discussões relevantes para classificar os autores, como a necessidade de se traduzir termos, já que alguns termos traduzidos podem perder seu significado, ou não ter um sentido na língua traduzida. Os autores utilizaram o índice de Jaccard para distribuir a similaridade entre os termos e aplicar a desambiguação.

Artigo 19: Desambiguação de nomes de autores para a identificação automática de perfis acadêmicos

O trabalho propõe uma estratégia de desambiguação para possibilitar a identificação automática do perfil de docentes do Google Acadêmico. Para tal, os autores realizam um casamento de nomes, comparando com as publicações cadastradas no currículo Lattes e no perfil do Google Acadêmico.

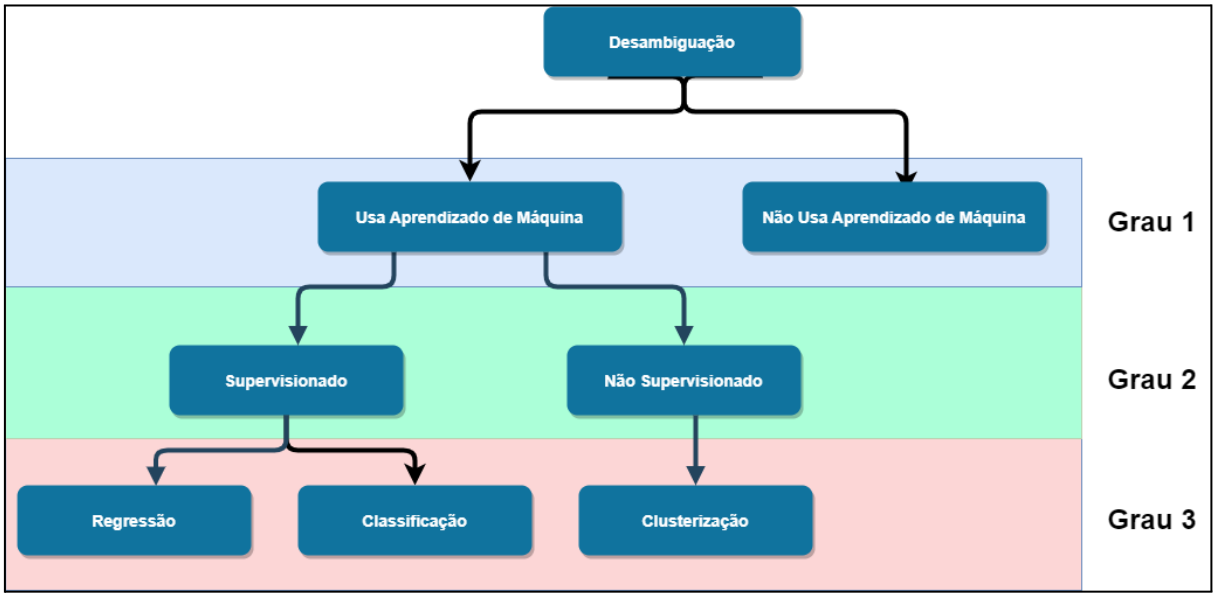
A decisão de desambiguação acontece, na validação desses perfis, procurando um perfil que apresente uma maior evidência de pertencer ao docente pesquisado.

5.1 Análise dos dados

A análise dos trabalhos separados levou a um modelo de classificação para representar o uso de aprendizado de máquina no processo de desambiguação de nome de autores. A classificação é demonstrada na figura 2. Para tal, três graus de classificação foram identificados, sendo eles:

- **Grau 1**: Segrega os trabalhos entre os que usam e os que não usam alguma técnica de aprendizado de máquina;
- **Grau 2**: Separa os trabalhos que usam técnicas de aprendizado de máquina entre aprendizado supervisionado e não supervisionado.
- **Grau 3**: Encapsula os trabalhos baseados na forma como os algoritmos utilizados funcionam, sendo elas regressão, classificação e agrupamento (clusterização); as duas primeiras formas são referentes a técnicas de aprendizado supervisionado, e a última, referente a técnicas de aprendizado não supervisionado.

Figura 2 - Uso de Aprendizado de Máquina para desambiguação



Fonte: o Autor (2022)

A maioria dos trabalhos analisados aplicam alguma técnica de aprendizado de máquina em várias partes do processo, seja na coleta, preparação e agrupamento dos dados ou mesmo nas análises dos resultados.

Ressalta-se que, mesmo aplicando técnicas de aprendizado supervisionado, alguns trabalhos aplicaram técnicas de aprendizado não supervisionado, mostrando que a utilização de uma não inviabiliza o uso do outra.

Quadro 2 - Trabalhos x Categorias

Artigo	Usa AM	Supervisionado / Não Supervisionado	Forma	Validação
Artigo 1	Sim	Supervisionado	Classificação	Manual
Artigo 2	Sim	Não Supervisionado	Agrupamento (Cluster)	Não listado
Artigo 3	Sim	Supervisionado Não Supervisionado	Classificação Agrupamento (Cluster)	Automática
Artigo 4	Sim	Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 5	Sim	Não Supervisionado	Agrupamento (Cluster)	Manual

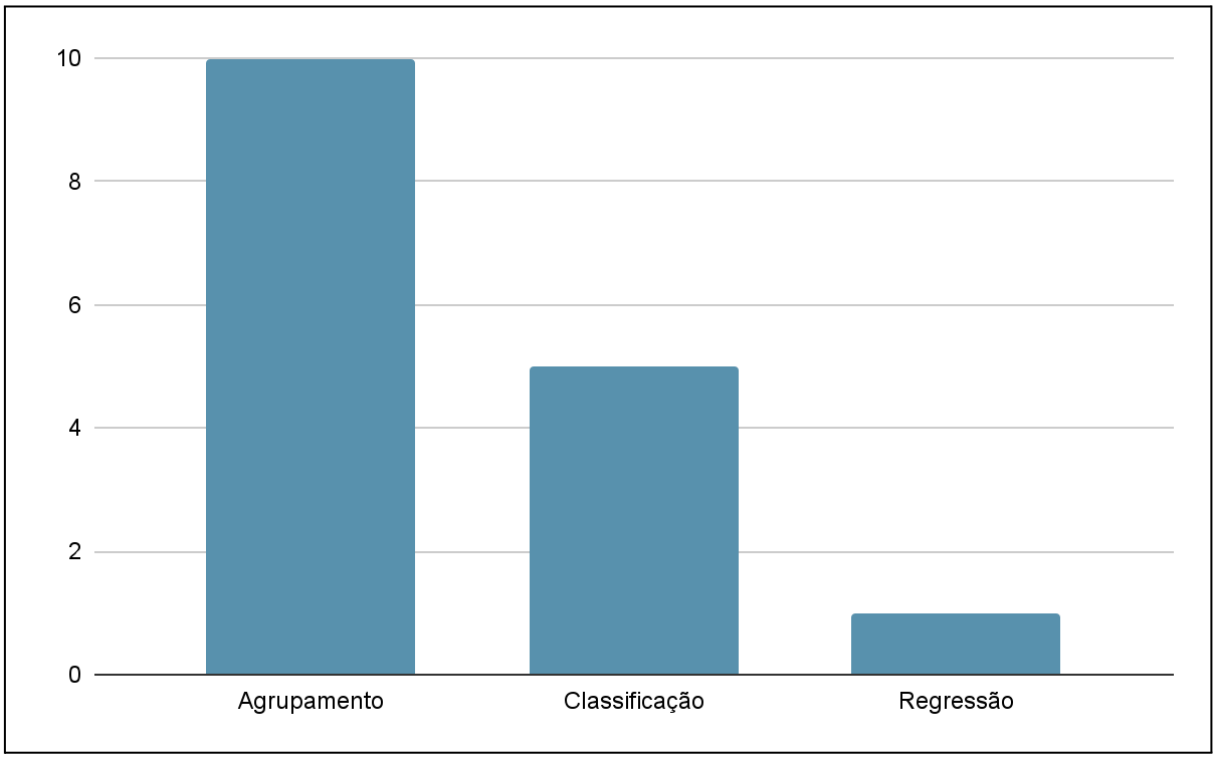
Artigo 6	Sim	Supervisionado	Classificação	Automática
Artigo 7	Não Usa			
Artigo 8	Sim	Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 9	Sim	Supervisionado	Classificação	Automática
Artigo 10	Sim	Supervisionado Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 11	Sim	Supervisionado	Regressão	Automática
Artigo 12	Sim	Supervisionado	Classificação	Automática
Artigo 13	Sim	Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 14	Sim	Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 15	Sim	Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 16	Sim	Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 17	Sim	Supervisionado	Classificação	Automática
Artigo 18	Sim	Não Supervisionado	Agrupamento (Cluster)	Automática
Artigo 19	Não utiliza			

Fonte: o Autor (2022)

A proporção entre uso e não uso do aprendizado de máquina é de 89.47% pró uso. Dos 19 trabalhos separados, apenas 2 não o utilizam, o que mostra uma tendência dos pesquisadores da área em utilizar algo do tipo para auxílio no processo de desambiguação de nomes de autores.

A figura 3 traz uma visão de como os trabalhos se dividem na aplicação, seja com aprendizado supervisionado ou não supervisionado.

Figura 3 - Distribuição de técnicas de aprendizado



Fonte: o Autor (2022)

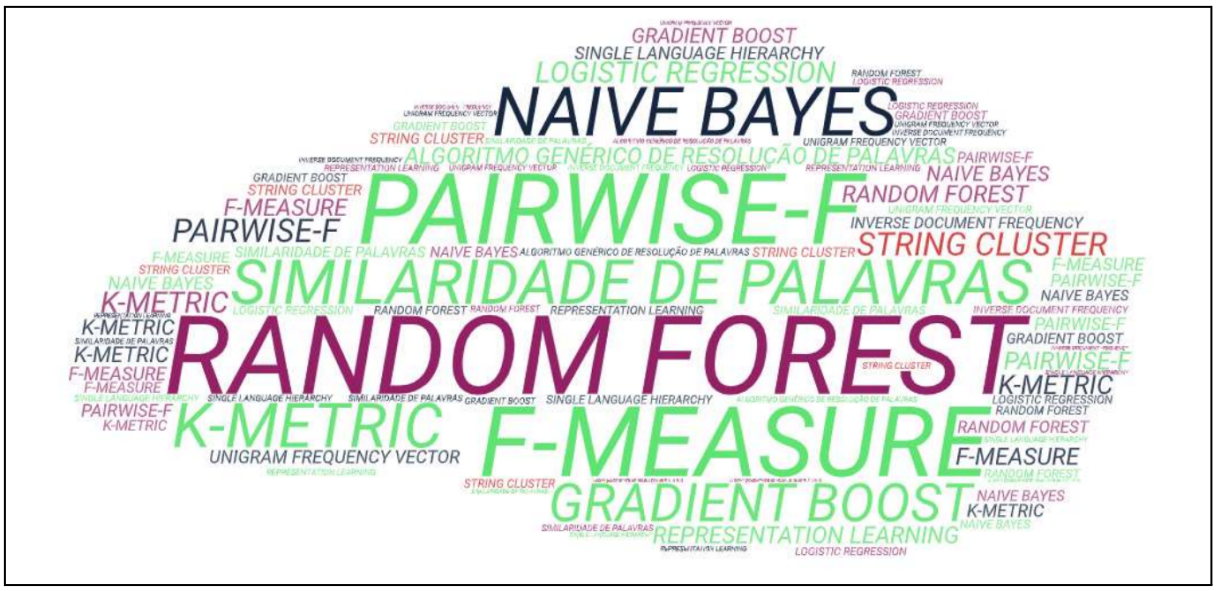
Os artigos levantados usam formas de agrupamento para várias partes do processo de desambiguação, seja para tratar os dados de origem, seja para gerar grupos de autores ou textos para representar o resultado separado.

Para agrupar os autores no geral, utilizam algoritmos com foco em buscar a semelhança de palavras, a também chamada “similaridade entre palavras”, ou grupos segregados, buscando a compatibilidade entre eles.

Algoritmos baseados em árvores (*Random-Forest, Pairwise-F, F-Measure*) aparecem com maior quantidade de aplicações/sugestões. Alguns trabalhos têm seu foco no âmbito teórico, explicitando um modelo a se seguir, mas não mostrando de fato como foi implementado.

Em suma, os trabalhos iniciam com uma base de dados coletados que, após passarem por diversas tratativas, são inseridos em processos de classificação ou agrupamento entre autores e trabalhos. O material coletado traz a visão de que a utilização de técnicas em conjuntos pode ser a forma mais efetiva de se realizar a desambiguação/unificação.

Figura 4 - Nuvem de palavras-chave de algoritmos de AM



Fonte: o Autor (2022)

A combinação dos filtros, que neste contexto são entidades com a capacidade de auxiliar na identificação única de um trabalho, ajuda na qualidade da desambiguação. Os filtros atuam como variáveis de agrupamento, sendo as mais utilizadas a rede de coautoria e o próprio nome do trabalho publicado.

Os trabalhos analisados somaram 13 filtros relevantes. O quadro 3, a seguir, apresenta essas categorias e o número de documentos representados por elas. Ressalta-se que um único texto pode ser representado por mais de uma categoria.

Quadro 3 - Categoria x Definição

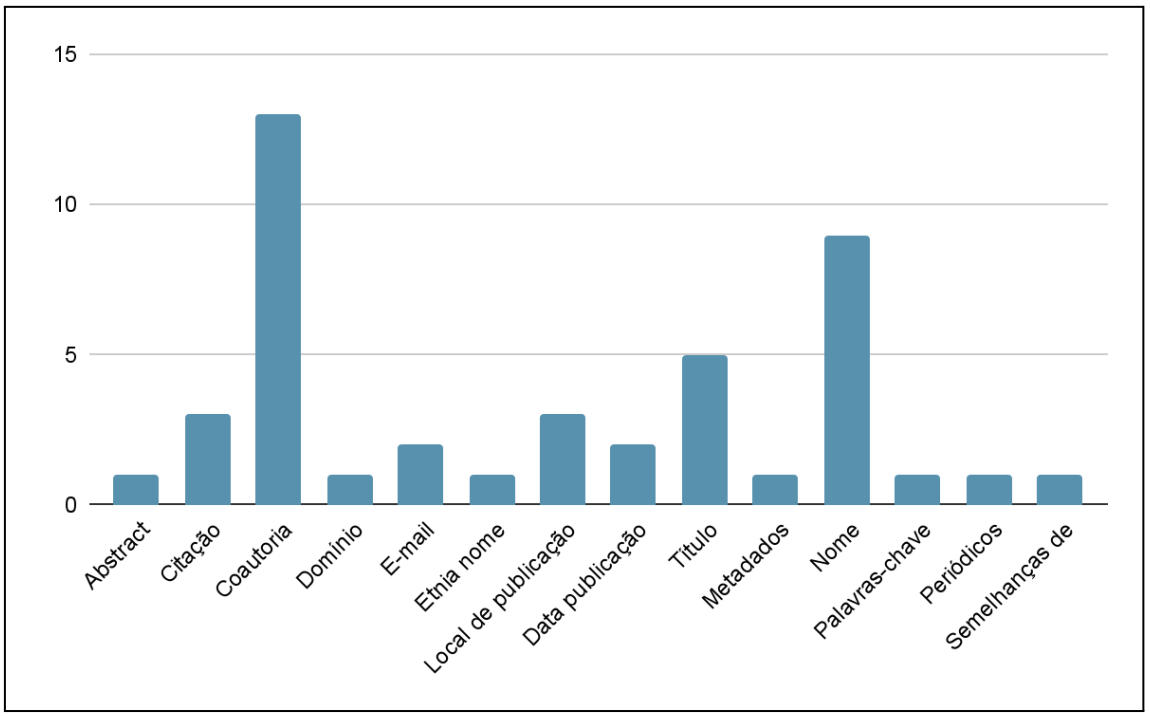
Categoria	Definição	Total
Abstract	Resumo do texto	1
Citação	Autores citados no trabalho	3
Coautoria	Coautores dos artigos	13
Domínio	Área de atuação do documento	1
E-mail	E-mail de contato do autor	2
Etnia nome	Similar ao nome, mas levando em consideração a etnia do mesmo	1
Local de publicação	Periódico, revista ou evento de publicação do trabalho	3

Data publicação	Data no qual o trabalho foi publicado	2
Título	Título do trabalho	5
Metadados	Dados intrínsecos ao documento	1
Nome	Nome do autor	9
Palavras-chave	Palavras-chave do trabalho	1
Periódicos	Revistas, Repositórios, Eventos	1
Semelhanças de palavras	Semelhança entre as palavras	1

Fonte: o Autor (2022)

Destacam-se rede de coautoria, nome do trabalho, título do trabalho e rede de citação como os filtros mais utilizados, com uma diferença considerável entre o primeiro citado e o último.

Figura 5 - Distribuição dos filtros



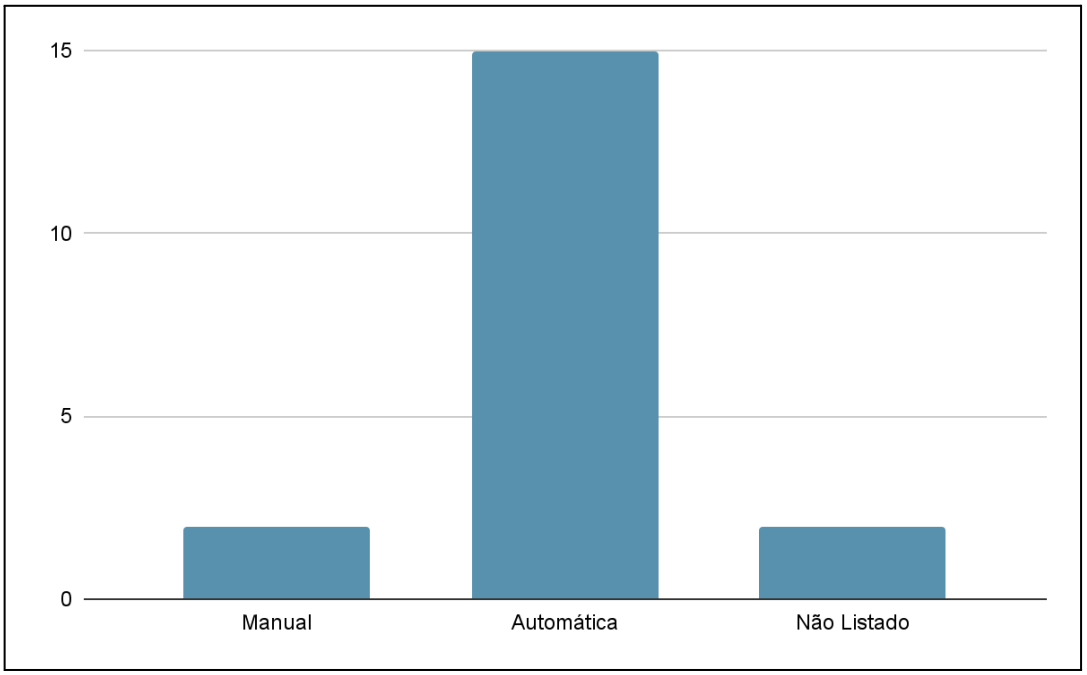
Fonte: o Autor (2022)

Para a aplicação, os trabalhos realizam a extração de informações para simular algum tipo de repositório de trabalhos científicos real. Alguns trabalhos retiram os dados de bases próprias como a PubMed, enquanto outros retiram de bases já estabelecidas, como a *Web of Science.*

Os artigos são multidisciplinares: há trabalhos relacionados à área da computação e da linguística, mas não há uma vertente focada na Ciência da Informação. Dessa forma, a desambiguação de nomes se apresenta como um meio, uma forma de seguir com o ciclo de vida dos dados e realizar atividades pertinentes de gestão e afins, não como um fim.

Como forma de avaliar o material já desambiguado, os trabalhos são quase unânimes na utilização de algoritmos de árvore (como o *Pairwise F*, por exemplo) para provar se a forma de desambiguação aplicada foi ou não efetiva.

Figura 6 - Distribuição das formas de avaliação



Fonte: o Autor (2022)

Dos artigos aceitos, apenas 2 (dois) não se referem à forma de validação do processo. Outros dois apresentam correção manual dos valores ao final do fluxo.

Dos trabalhos que aplicam uma conferência automática dos resultados, destaca-se que alguns utilizaram *datasets* com referências desambiguadas para auxílio na conferência final. Entretanto, como a validação é realizada de forma automática, esses trabalhos se mantiveram nessa categoria, ao invés de se criar uma nova categoria para trabalhos validados de forma “semi-automática”.

Esse processo é melhor discutido na próxima seção, “Fluxo de Desambiguação de Nome de Autores”, que apresenta um modelo de desambiguação junto a explicações mais aprofundadas sobre cada etapa.

Em resumo, este mapeamento sistemático da literatura mostrou que a desambiguação de nomes de autores é uma tarefa complexa que tem sido abordada em diferentes áreas, principalmente em computação e linguística. A maioria dos artigos analisados utilizou algoritmos de agrupamento baseados em similaridade entre palavras para classificar e agrupar autores e trabalhos, combinando diversos filtros para aumentar a efetividade da desambiguação. Os filtros mais utilizados foram a rede de coautoria, nome do trabalho, título do trabalho e rede de citação. A avaliação dos resultados geralmente foi feita por meio de algoritmos de árvore e alguns trabalhos incluíram uma correção manual dos valores ao final do processo. O modelo de fluxo de desambiguação de nome de autores apresentado na próxima seção se baseia nessas conclusões e fornece uma estrutura mais detalhada para a realização dessa tarefa desafiadora.

6. FLUXO DE DESAMBIGUAÇÃO DE NOME DE AUTORES

Com base no material levantado e analisado pelo MSL, pode-se observar um fluxo quase obrigatório na implementação de qualquer método de desambiguação da informação de entidades, passando pela construção da base de dados, pela desambiguação das informações e pela validação do resultado obtido.

O processo, por definição, precisa de uma estrutura de dados-base para iniciar a aplicação. Essa estrutura, que pode ser tratada ou não, neste contexto será chamada de "*dataset*".

A construção do *dataset* inicial, na grande maioria das vezes, é realizada a partir de um fragmento que representa uma base de dados robusta, na qual se obtém material mapeado e com ambiguações conhecidas para facilitar o processo de validação.

Wu e Ding (2012) utilizaram a "ISI WoS dataset" como ponto de partida para construção de sua base, continuando com uma "busca avançada" no próprio WoS, buscando artigos das áreas de ‘‘Information Science & Library Science’’. Todo o conteúdo encontrado foi salvo.

Gurney, Horlings e Besselaar (2012) também recorreram ao WoS para construção da base inicial, analisando o material coletado com uma ferramenta citada em outro trabalho (SAINT, Somers *et al.*, 2009), escolhendo manualmente autores que mantinham mais de uma publicação e sobrenome compartilhado com outros escritores. Wang *et al* (2012) seguem esta fórmula, construindo um cenário menor com um conjunto de 4253 artigos pertencentes a 100 autores americanos distintos, com o objetivo de facilitar o processo de análise de resultado de cada artigo individualmente. Pooja, Mondal e Chandra (2020) utilizaram o *dataset* de *"Arnetminer name"*, enquanto Hussain e Asghar (2018) recorreram aos *datasets* de "*Arnetminer-S*" e "*Arnetminer-L*" em seus respectivos processos, já que os conjuntos de dados apresentados já têm conteúdo avaliado e classificado, apresentando material já desambiguação, agilizando o processo de conferência.

Trabalhos que atuam com essa ramificação, obtendo dados de fontes externas ou de fontes já estabelecidas, trazem um viés mais focado em testar a fórmula de desambiguação apresentada, não necessariamente aplicando a fórmula em questão como melhoraria e/ou solução de algum problema prático de algum

ambiente digital científico. Assim, acabam gerando uma problemática em seu processo de aplicação geral, no qual os *datasets* precisam passar por um processo de transformação para se adequar ao método, ou vice-e-versa.

A aplicação em cenários específicos precisaria se atentar a quais filtros são pertinentes nos métodos que utilizam tais *datasets* para compor a base documental correta. Isso é necessário para não afetar o resultado final, levando ao fluxo de "Obtenção de dados relevantes", no qual dados de fontes externas são consultados e integrados ao *dataset* inicial, reiniciando o fluxo e a aplicação do método de desambiguação.

Wang et Al (2012) optaram por utilizar registros da WoS para compor os metadados do *dataset* inicial, como nome dos autores, trabalhos citados, palavras-chave combinadas, título, *stop words* e categorias, utilizando informações como nome e rede de afiliação como buscas principais para obtenção de textos relevantes. Neste caso, a busca por mais material documental se fez necessária pelo fato de o *dataset* não ter informações iniciais suficientes para passar por um processo de desambiguação, sendo a consulta de meios externos um apoio para melhoria do processo.

Após a construção do *dataset* inicial, ou atualização desse *dataset* por meio de fontes externas de uma versão previamente gerada, o fluxo segue para a aplicação do método de desambiguação em si, no qual o método utilizado demanda seus filtros para atuação em seu desenvolvimento como um todo.

Antes do encerramento do processo, a "Desambiguação" pode ser refeita para garantir refino, antes de evoluir para a fase de "Validação", na qual as informações desambiguadas são avaliadas – o que leva ao fim do fluxo.

Pode-se generalizar e dizer que os métodos de desambiguação que utilizam classificação, regressão e clusterização podem gerar valores classificativos como palavras com erros ortográficos, variantes de mesmo nome ou casos afins. Neste contexto, Carrasco et al (2016) afirmam que:

> expressões temporais combinadas com medidas de similaridade devem fornecer grupos de nomes que são variantes ou erros de digitação da forma preferida. Uma vez que apenas a inspeção manual pode determinar a variante preferida e alguns casos difíceis precisam sempre ser revisados por especialistas[…].

Entretanto, os valores de classificação/agrupamento podem passar por diversos processos de refino, como a utilização de Processamento de Linguagem Natural (PLN), diminuindo “ruídos” informacionais gerados por estes processos.

Ainda em contrapartida à afirmação de Carrasco et al (2016), a grande maioria dos trabalhos separados pelo MSL trouxeram validações de forma automática, provando que a possibilidade do uso de técnicas de aprendizado de máquina nesta etapa é bem quista pela comunidade acadêmica.

Zhu et al (2016) utilizam apenas a rede de coautoria para conduzir o processo de clusterização, por considerarem-na forte o suficiente para garantir que o processo seja realizado, conseguindo um resultado de aproximadamente 100% de aproveitamento na validação com *PairwisePrecision*.

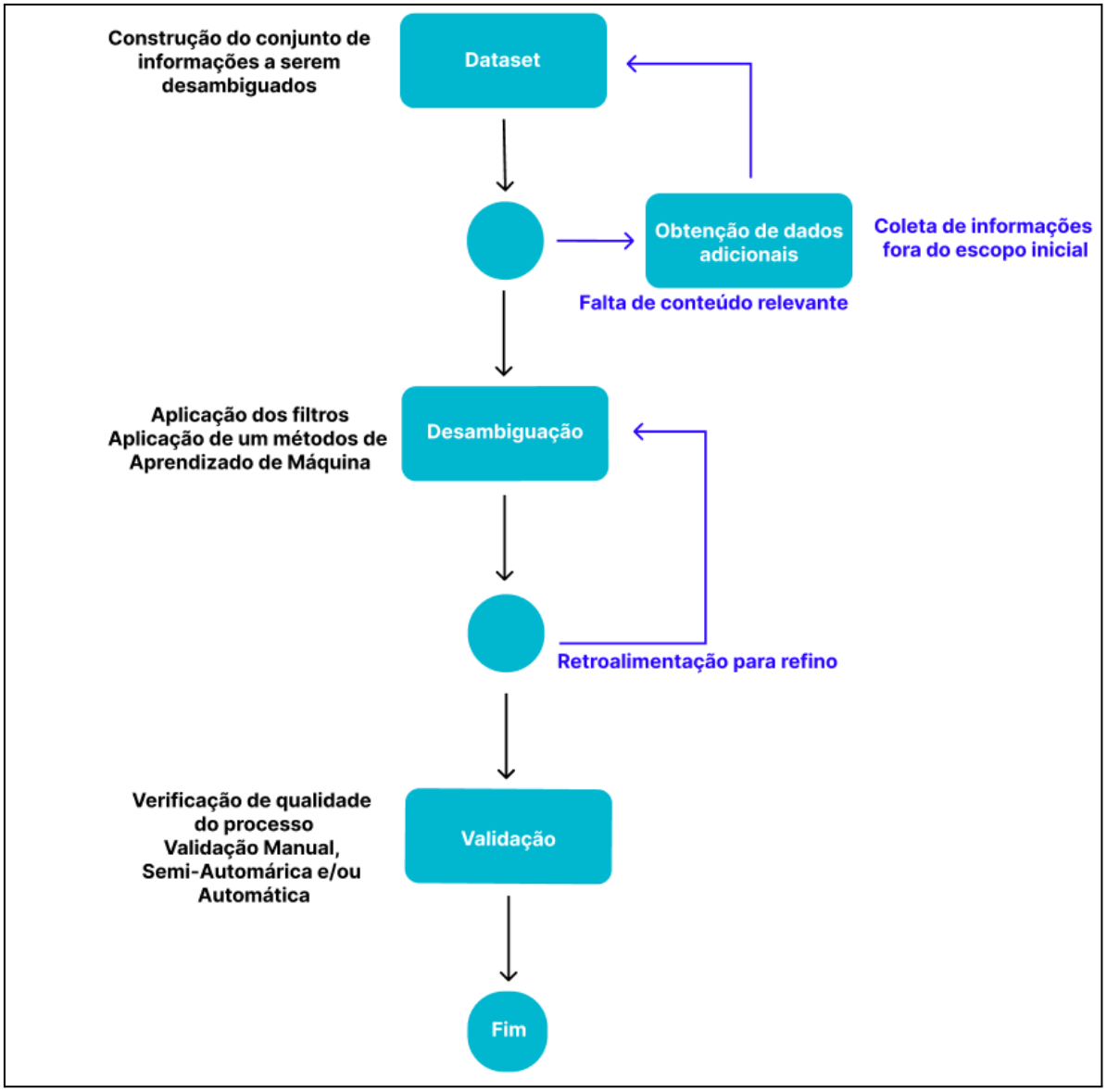
Gurney, Horlings e Besselaar (2012) calculam a precisão utilizando o *F-measure* dos autores e a contagem de publicação, baseando-se no último nome e na primeira inicial, assim como Mihaljević e Santamaria (2021), Kim, Kim e Owen-Smith (2022) e outros trabalhos aprovados pelo mapeamento.

A partir desse contexto, o processo de desambiguação automática pode ser ajustado de acordo com a necessidade de preferência de seu construtor, mas a literatura aponta preferência pela utilização de algoritmos de aprendizado de máquina não-supervisionados.

Dentro do contexto apresentado, algumas etapas foram elencadas foram seguidas quase que por padrão pela maioria dos artigos, sendo elas a construção do *dataset* e a fase de desambiguação e validação dos resultados, gerando um fluxo simplificado. Porém, quando analisado por etapas, este fluxo pode gerar muitas ramificações de estudo e funcionamento, ou levar a outras etapas “opcionais”.

Dito isso, o fluxograma de desambiguação apresentado na figura 7 foi gerado, levando em consideração apenas o conteúdo recuperado pelo MSL.

Figura 7 - Fluxograma da desambiguação



Fonte: o Autor (2022)

A modelagem do *dataset*, em sua essência, apresenta apenas uma aglomeração do banco de dados do ADC. Uma vez que este banco é separado, algumas opções podem ser seguidas, como a segregação em pacotes informacionais menores, levando a desambiguações mais rápidas que resultam em um processo de validação mais facilitado, podendo ser feito de forma automática ou manual. Todavia, separar as bases em escopos menores pode levar à necessidade de uma nova avaliação, já que trabalhos de um autor ou instituição podem estar anexados em blocos diferentes, tornando uma análise do conteúdo total armazenado no ADC uma solução mais assertiva.

Uma vez que a base não apresenta informações o suficiente para executar a fase de desambiguação de forma mais satisfatória, recursos podem ser aplicados para se coletar mais conteúdo em sites institucionais, bases de pesquisa e afins, remodelando o *dataset* inicial.

A fase de desambiguação trabalha com a informação disponibilizada, tratando-a para que possa atuar no algoritmo selecionado, com a opção de repetir o processo diversas vezes, levando à fase final de validação, na qual os resultados são analisados de forma manual ou automática para garantir o sucesso da operação.

O modelo foi gerado levando-se em consideração apenas os trabalhos extraídos pelo mapeamento, entendendo e comparando o funcionamento dos mesmos. Isto posto, os fluxos podem ser destrinchados da seguinte forma:

Fase 1 - *Dataset*:

Constituída pelo levantamento do material a ser desambiguado, reflete uma parte dos dados de uma base ou o seu material na íntegra, podendo o levantamento ser complementado pelo material coletado na fase 2.

Com o algoritmo de desambiguação bem definido, o *dataset* pode remover informações não pertinentes ao processo, diminuindo a quantidade de informação a ser analisada, o que por sua vez diminui o tempo de execução do fluxo. Caso o algoritmo não esteja bem definido, a separação do material na íntegra segue como opção principal.

Fase 2 - Obtenção de dados adicionais:

Fase opcional, executada somente quando constatada a deficiência informacional do *dataset*, e que tem a finalidade exclusiva de recuperar informação faltante para completar o quadro de dados a respeito dos trabalhos.

Podem-se utilizar dados de autoridades *web scrappers* ou qualquer outra forma de se obter informação necessária. Os *scrappers* são a solução mais utilizada, e consistem em robôs programados para retirar informações de páginas na internet. Páginas de revistas e acervos (institucionais ou mesmo pessoais) podem ser utilizados para coletar esse tipo de dado. A utilização dos dados de autoridade

pode ser uma opção viável, sendo trazida apenas como uma possibilidade, já que não foi explorada dentro do contexto deste trabalho.

Fase 3 - Desambiguação da Informação:

A fase de desambiguação é a principal parte do fluxo, pois é responsável por aplicar técnicas à informação selecionada no *dataset*, com o objetivo de separar informações duplicadas e unificar dados do mesmo contexto.

Para isso, as entidades passam por processos de limpeza, adequando-se ao modelo de aprendizado de máquina proposto. São realizados fluxos de processamento de linguagem natural complexos, como a remoção de palavras sem sentido, tais como “em”, “o”, “as”, “a” e afins, ou outras aplicações ligadas à área, eliminando conteúdo não relevante para o algoritmo e limpando a base.

Caso o filtro escolhido seja apenas a rede de coautoria, qualquer informação extra pode ser descartada, e apenas conteúdos que referenciam esta temática seriam selecionados, funcionando da mesma forma independente do filtro selecionado.

Outras implementações podem ser necessárias, para preparar as informações para o algoritmo a ser utilizado, deixando aberta a possibilidade de uso de mais de um algoritmo no fluxo, detalhe recomendado por muitos trabalhos.

Tanto para a utilização de algoritmos supervisionados quanto para não supervisionados, a abordagem de desambiguação por blocagem será a mais utilizada, separando as entidades em blocos menores para serem classificados ou agrupados. O resultado desse processo são os próprios blocos, onde cada um vai representar uma entidade unificada. Recomenda-se o uso de algoritmos não supervisionados, pois obtiveram um resultado melhor em comparação aos supervisionados.

Pode-se exemplificar a desambiguação utilizando o algoritmo de *Naive Bayes*, gerando uma tabela de probabilidade pós-classificação dos dados iniciais para treinamento e a desambiguação utilizando o algoritmo de *Random Forest*, onde os trabalhos atuariam como concorrentes em cada nó, que verificaria o grau de similaridade, construindo uma árvore de decisão.

A forma de desenvolvimento de ambos os algoritmos é divergente, mas os blocos classificados e separados se mantêm como resposta válida, sejam eles

representados pela própria tabela de probabilidade ou por uma árvore de decisão.

O funcionamento do algoritmo se restringe diretamente a esta etapa, já que ela tem o objetivo de entregar a etapa posterior a informação desambiguada, independente do método ou forma utilizados para tal. Dessa forma, ajustes podem ser realizados para um melhor funcionamento da desambiguação a partir das peculiaridades de cada algoritmo selecionado.

Contida nessa fase está o processo de **Retroalimentação para refino**, sendo apenas novas aplicações dos métodos já aplicados, mas utilizando o resultado obtido ao final do ciclo como o *dataset* inicial. Como os dados já foram tratados e segmentados em blocos, a fase de refino confirmaria os blocos existentes, removeria blocos desnecessários e melhoraria o agrupamento, trazendo um pouco mais de confiabilidade aos resultados.

As novas aplicações podem trazer uma capacidade de aprimoramento no resultado final, já que blocos menores vão estar disponíveis para agrupamento. Todavia, a retroalimentação pode ocasionar um fluxo mais extenso e demorado.

Fase 4 - Validação:

A última fase, validação, lida com a verificação da qualidade do fluxo anterior, tentando encontrar possíveis divergências. Algoritmos de aprendizado não supervisionado são uma opção viável para esta etapa, podendo ser a única linha de validação ou ser acompanhada por validação manual, de modo a garantir acurácia maior ao fluxo. Para essa fase, algoritmos que trabalham medindo a precisão e *recall* são empregados, obtendo como “produto” uma porcentagem de acerto.

Ressalta-se que esta fase é abordada nos trabalhos listados do mapeamento, em sua maioria, como uma prova de conceito do método analisado, fato que não exclui sua utilização em cenários reais.

O fluxo de desambiguação automática da informação apresentando se mostra relativamente simples, com fases cujos papéis são bem definidos e distintos. Entretanto, mesmo que o fluxo seja direto, cada fase apresenta suas peculiaridades e dificuldades, sendo necessário o apoio entre profissionais de áreas distintas para que o fluxo seja implementado de forma consciente e consistente.

7. CONSIDERAÇÕES FINAIS

Este trabalho teve o objetivo de identificar as formas com as quais os pesquisadores realizam a desambiguação automática de entidades utilizando técnicas de aprendizado de máquina, detalhando como o processo é realizado de forma geral, bem como suas especificações.

Para atingir este objetivo geral, seguiu-se rumo ao objetivo específico de “contextualizar a respeito da desambiguação da informação e aprendizado de máquina”. Este objetivo foi contemplado na Seção 3, na qual tais temas foram abordados. O estudo bibliográfico permitiu compreender o que é a área de desambiguação da informação, em conjunto com o funcionamento do aprendizado de máquina, e elaborar uma breve descrição de alguns algoritmos levantados pelo mapeamento que embasa o trabalho.

O objetivo específico de “realizar um mapeamento sistemático da literatura (MSL), para identificar modelos de desambiguação automática de nome de autores”, foi contemplado na Seção 4. O MSL permitiu compreender como a literatura tem abordado a desambiguação de nome de autores e instituições. Os 19 artigos selecionados puderam ser divididos em três categorias, que exemplificam o processo de desambiguação automática – as quais foram exploradas na Seção 5, construída considerando a base documental levantada pelo mapeamento.

A primeira categoria selecionada se refere ao uso de AM no processo de ambiguação, identificando que a maioria dos trabalhos listados aderem a utilização de técnicas de aprendizado não-supervisionado para desambiguar. Outra categoria tratou da forma com a qual os artigos avaliaram seus resultados, podendo ser de forma automática ou manual.

Demonstrou-se uma preferência pela validação automática do material desambiguado. Em outras palavras, a análise resultou na constatação de uma preferência pela utilização de técnicas de aprendizado de máquina para a desambiguação da informação e no processo de validação dos dados desambiguados, processo que na maioria das vezes também é realizado de forma automática. Existe também uma preferência pelo uso de algoritmos de aprendizado não-supervisionado em ambos os processos.

Por fim, a última categoria separada se refere aos filtros utilizados pelo fluxo de desambiguação. Foi possível verificar que a realização da desambiguação das

entidades está diretamente ligada ao processo de estipulação dos “filtros”, elementos textuais que compõem os trabalhos e podem ser utilizados para segregar um artigo de outro, sendo a rede de coautoria e o título do trabalho os filtros mais utilizados, o que não inviabiliza a combinação de vários filtros para melhorar a acurácia da desambiguação.

A partir dessas análises, foi construída a Seção 6, propondo um modelo para exemplificar o processo de desambiguação, detalhando cada etapa e possíveis variações existentes no processo, seguindo esta ordem:

- A construção do *dataset*;
- Obtenção de dados adicionais;
- Desambiguação;
- Retroalimentação para refino;
- Validação.

Ressalta-se, como desafio primordial, garantir a homogeneidade dos dados, já que as bases reais podem não conter o conteúdo informacional necessário para executar o processo de desambiguação. Essa é uma problemática não enfatizada por todos os trabalhos, devido ao uso das bases de dados mais apropriadas para avaliar diretamente o método utilizado. Para tanto, os processos de construção de *dataset* e a obtenção de dados adicionais estão envolvidos, garantindo os dados faltantes de formas não mapeadas.

A falta de informações relevantes dentro de uma base de dados específica pode induzir qualquer processo de desambiguação ao erro, mostrando que bases com uma quantidade de dados mais abrangente podem facilitar essa separação ou unificação. Para que outros dados sejam anexados ao *dataset*, robôs podem ser programados para buscar dados dos autores e trabalhos em portais acadêmicos, páginas pessoais ou acervos digitais.

Para a desambiguação, algoritmos de aprendizado não supervisionado mostraram desempenho melhor, atuando a partir de técnicas de blocagem para realizar a desambiguação de entidades.

Novas inserções no método escolhido para desambiguar podem ser realizadas, a combinação entre métodos sendo inclusive recomendada, respondendo como “Retroalimentação e refino”.

Para a fase final de “Validação”, algoritmos de aprendizado não supervisionado também se mostraram uma opção viável, trabalhando com cálculos de precisão e *recall* com um resultado mais estatístico da desambiguação, podendo ser realizada em conjunto com a validação manual, que não é descartada.

Após detalhar os resultados da pesquisa, pode-se entender que o objetivo geral foi atendido, já que a forma com a qual a literatura está abordando a desambiguação da informação de entidades de forma automática foi detalhada e explorada. Assim, acredita-se que este trabalho possa contribuir como um ponto de partida para novos materiais no campo da desambiguação da informação, pois indica como o fluxo total é realizado e validado, fornecendo exemplos de como outros autores trabalharam dentro do fluxo e mapeando o processo como um todo. Este trabalho traz premissas que podem ser estudadas futuramente, ampliando o conhecimento a respeito da área, garantindo melhores resultados no processo.

Para pesquisas futuras, propõe-se um aprofundamento maior sobre o uso de técnicas de aprendizado de máquina combinadas com formas diversas de agrupamento, aplicando-as em algum cenário hipotético para validação.

Abre-se a sugestão de se trabalhar com elementos menores descobertos ao longo do trabalho, como formas de se gerar as redes ligadas aos filtros (como, por exemplo, a rede de coautoria) de forma automática e retroalimentada.

Outra vertente apresentada seria uma abordagem mais minuciosa, ligando os dados de autoridade às etapas apresentadas no fluxo de desambiguação automática gerado por este trabalho, trazendo maior ligação entre as temáticas.

Acrescenta-se que este trabalho contribui para o cenário científico servindo como um norte para auxiliar trabalhos relacionados a desambiguação da informação utilizando inteligência artificial. No cenário social, a aplicação desmistificada pode facilitar a aplicação em ambientes como o ENANCIB, permitindo a extração de métricas úteis para tais ambientes.

REFERÊNCIAS

ALVES, Rachel Cristina Vesú**. Metadados como elementos do processo de catalogação**. 2010. 132 f. Tese (Doutorado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em:https://repositorio.unesp.br/bitstream/handle/11449/103361/alves_rcv_dr_mar.pdf . Acesso em: 03 jan. 2023.

AMBIGUAÇÃO. **In: DICIO, Dicionário Online de Português.** Porto: 7Graus, 2022. Disponível em: https://www.dicio.com.br/ambiguacao. Acesso em: jul. 2022.

BATISTA, Gustavo Enrique de Almeida Prado Alves. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/en.php. Acesso em: jul. 2022.

BIANCHI, Reinaldo Augusto da Costa. **Uso de Heuristicas para a Aceleração do Aprendizado por Reforço.** Disponível em: https://www.teses.usp.br/teses/disponiveis/3/3141/tde-28062005-191041/publico/tese-bianchi.pdf. Acesso em: abr. 2022.

BRASCHER, Marisa. **A ambiqüidade na recuperação da informação**. Disponível em: https://ridi.ibict.br/handle/123456789/284. Acesso em: jun. 2022.

BONTHU, Harika. **Single-Link Hierarchical Clustering Clearly Explained!**. Disponível em: https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/. Acesso em: jul. 2022.

CARRASCO, Rafael C. SERRANOA, Aureo. CASTILLO-BUERGO, Reydi. **A parser for authority control of author names in bibliographic records**. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0306457316300085. Acesso em: mar. 2022

CERRI, Ricardo. CARVALHO, André Carlos Ponce de Leon Ferreira de. **Aprendizado de Máquina:** Breve Introdução e Aplicações. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/184785/1/Aprendizado-de-maquina-breve-introducao.pdf. Acesso em: mar. 2022.

CFHSS OCS Guide. Disponível em: https://pkp.sfu.ca/files/FedCan_OCS_Guide.pdf. Acesso em: set. 2022.

CLACK, D. H. Authority control and linked bibliographic databases. Cataloging & Classification Quarterly, [S.L], v. 8, n. 3/4, p. 35-46, 1988. Disponível em: < https://www.tandfonline.com/doi/abs/10.1300/J104v08n03_03 >. Acesso em: 04 jan. 2023.

DEMBCZYNSKI, Krzysztof. WAEGEMAN, Willem. CHENG, Weiwei. HÜLLERMEIER, Eyke. **An Exact Algorithm for F-Measure Maximization**. Disponível em: https://proceedings.neurips.cc/paper/2011/file/71ad16ad2c4d81f348082ff6c4b20768-Paper.pdf. Acesso em: jan. 2023.

DESAMBIGUAÇÃO. **In: DICIO, Dicionário Online de Português.** Porto: 7Graus, 2022. Disponível em: https://www.dicio.com.br/desambiguacao. Acesso em: jul. 2022.

DIGIAMPIETRI, L. A.; FERREIRA, J. E. Desambiguação de nomes de autores para a identificação automática de perfis acadêmicos. **Em Questão,** v. 24, n. 2, p. 37-54, 2018. DOI: 10.19132/1808-5245242.37-54.

DING, Xiu-Hao. WU, Jiang. **Author name disambiguation in scientific collaboration and mobility cases.** Disponível em: https://dl.acm.org/doi/10.1007/s11192-013-0978-8. Acesso em: mar. 2022.

Fluxo da Informação. **Quando usar F1 Score?**. Disponível em: https://fluxodeinformacao.com/biblioteca/artigo/read/57163-quando-usar-f1-score; Acesso em: set. 2022.

GURNEY, Thomas. HORLINGS, Edwin. BESSELAAR, Peter van den. **Author disambiguation using multi-aspect similarity indicators.** Disponível em: https://www.researchgate.net/publication/223966499_Author_disambiguation_using_multi-aspect_similarity_indicators. Acesso em: mar. 2022.

GUSMÃO, Amanda. **O que é Naive Bayes e como funciona esse algoritmo de classificação**. Disponível em: https://rockcontent.com/br/blog/naive-bayes/#:~:text=Naive%20Bayes%20é%20um%20algoritmo,no%20meio%20acadêmico%20da%20estatística. Acesso em: set. 2022.

HUANG, Shuiqing. YANG, Bo. YAN, Sulan. ROUSSEAU, Ronald. **Institution name disambiguation for research assessment.**

HUSSAIN, Ijaz. ASGHAR, Sohail. **Disambiguating homonyms using graph structural clustering.**

ICMS Júnior. **O que é Random Forest?.** Disponível em: https://icmcjunior.com.br/random-forest/#:~:text=O%20que%20é%20Random%20Forest,para%20chegar%20no%20resultado%20final. Acesso em: set. 2022.

JESUS, Ananda Fernanda. **RECOMENDAÇÕES TEÓRICO-METODOLÓGICAS PARA A PUBLICAÇÃO DE DADOS BIBLIOGRÁFICOS ABERTOS E CONECTADOS**. Disponível em: https://repositorio.ufscar.br/handle/ufscar/14228?show=full. Acesso em: 03 jan. 2023.

KIM, Jinseck. KIM, Jinmo. OWEN-SMITH, Jason. **Generating automatically labeled data for author name disambiguation:** an iterative clustering method. Disponível

em: https://link.springer.com/article/10.1007/s11192-018-2968-3. Acesso em: mar. 2022.

KIM, Jinseok. **Evaluating author name disambiguation for digital libraries**: a case of DBLP. Disponível em: https://link.springer.com/article/10.1007/s11192-018-2824-5. Acesso em: mar. 2022.

KIM, Jinseok. KIM, Jenna. OWEN-SMITH Jason. **Ethnicity-based name partitioning for author name disambiguation using supervised machine learning**. Disponível em: https://ideas.repec.org/a/bla/jinfst/v72y2021i8p979-994.html. Acesso em: mar. 2022.

KISHIDA, Kazuaki. **Term disambiguation techniques based on target document collection for cross-language information retrieval:** an empirical comparison of performance between techniques. Disponível em: https://dl.acm.org/doi/10.1016/j.ipm.2006.04.006. Acesso em: mar. 2022.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering: a tertiary study**. Information and Software Technology,** [S.l.], v. 52, p. 792–805, 2010.

KM, Pooja. MONDAL, Samrat. CHANDRA, Joydeep. **A Graph Combination With Edge Pruning-Based Approach for Author Name Disambiguation.** Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24212. Acesso em: mar. 2022.

KIDO, Guilherme Sakaji. BARBON JUNIOR, Sylvio. MORIGUCHI, Stella Naomi. **Comparação entre TF-IDF e LSI para pesagem de termos em micro-blog**. Disponível em: https://www.researchgate.net/publication/262804841_Comparacao_entre_TF-IDF_e_LSI_para_pesagem_de_termos_em_micro-blog. Acesso em: nov/2022.

LEVIN, Felipe Hopper**. Desambiguação de Autores em Bibliotecas Digitais utilizando Redes Sociais e Programação Genética.** Disponível em: https://www.lume.ufrgs.br/handle/10183/26968. Acesso em: mar. 2022.

LIU, Wanli. DOGAN, Rezarta Islamaj. KIM, Sun. COMEAU, Donald C. KIM, Won. YEGANOVA, Lana. LU, Zhyong. **Author Name Disambiguation for PubMed.** Disponível em: https://pubmed.ncbi.nlm.nih.gov/28758138/. Acesso em: mar. 2022.

LIU, Yu. LI, Weijia. HUANG, Zhen. FANG, Qiang. **A Fast Method Based on Multiple Clustering for Name Disambiguation in Bibliographic Citations.** Disponível em:https://www.researchgate.net/publication/264759553_A_Fast_Method_Based_on_Multiple_Clustering_for_Name_Disambiguation_in_Bibliographic_Citations . Acesso em: mar. 2022.

MACULAN, Benildes Coura Moreira dos Santos. AGANETTE, Elisângela Cristina. **Desambiguação de relações em tesauros e o seu reúso em ontologias.**

BRAPCI, 2014. Disponível em: https://brapci.inf.br/index.php/res/v/19063. Acesso em: mar. 2022.

MARAUT, Stéphane, MARTÍNEZ, Catalina. **Identifying author–inventors from Spain: methods and a first insight into results.** Disponível em: https://link.springer.com/article/10.1007/s11192-014-1409-1. Acesso em: mar. 2022.

MIHALJEVIC, Helena.SANTAMARIA, Lucía**. Disambiguation of author entities in ADS using supervised learning and graph theory methods.** Disponível em: https://link.springer.com/article/10.1007/s11192-021-03951-w. Acesso em: mar. 2022.

MODARD, Maria Carolina. Baranauskas, José Augusto. **Conceitos sobre Aprendizado de Máquina.** Disponível em: https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf. Acesso em: abr. 2022

MONZ, Christof, WEERKAMP, Wouter. **A Comparison of Retrieval-Based Hierarchical Clustering Approaches to Person Name Disambiguation**. Disponível em: https://www.researchgate.net/publication/221300463_A_Comparison_of_Retrieval-Based_Hierarchical_Clustering_Approaches_to_Person_Name_Disambiguation. Acesso em: maio. 2022

NICK, Todd G**.** CAMPBELL**,** Lathleen M**. Logistic Regression.** Disponível em: https://link.springer.com/protocol/10.1007/978-1-59745-530-5_14. Acesso em nov/2022.

OPEN JOURNAL SYSTEMS. Disponível em: https://pkp.sfu.ca/ojs/. Acesso em: set. 2022.

PopGen.jl. **Pairwise F-Statistics**. Disponível em: https://biojulia.net/PopGen.jl/docs/analyses/fstatistics. Acesso em: set. 2022.

Portal do Bibliotecário. **O Que É DSpace**. Disponível em: https://portaldobibliotecario.com/biblioteconomia/o-que-e-dspace/index.html. Acesso em: set. 2022

SANCHES, Marcelo Kamiski. **Aprendizado de máquina semi-supervisionado:** proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados**.** Disponível em: https://teses.usp.br/teses/disponiveis/55/55134/tde-12102003-140536/publico/Dissertacao_MKS.pdf. Acesso em: abr. 2022

SANTANA, Alan Filipe. GONÇALVES, Andre. LAENDER, Alberto H. F. FERREIRA, Anderson A. **Incremental Author Name Disambiguation by Exploiting Domain-Specific Heuristics**. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23726. Acesso em: mar. 2022.

SANYAL, Kumar Debarshi. BHOWMICK, Plaban Kumar. DAS, Partha Pratim. **A review of author name disambiguation techniques for the PubMed bibliographic database.** Disponível em: https://journals.sagepub.com/doi/abs/10.1177/0165551519888605?journalCode=jisb. Acesso em: mar. 2022.

SHIN, Dongwook. KIM, Taehwan. CHOI, Joongmin. Kim, Jungsun. **Author name disambiguation using a graph model with node splitting and merging based on bibliographic information.** Disponível em: https://link.springer.com/article/10.1007/s11192-014-1289-4. Acesso em: mar. 2022.

SILVA, Johny. **Uma breve introdução ao algoritmo de Machine Learning Gradient Boosting utilizando a biblioteca Scikit-Learn**. Disponível em: https://medium.com/equals-lab/uma-breve-introdução-ao-algoritmo-de-machine-learning-gradient-boosting-utilizando-a-biblioteca-311285783099#:~:text=O%20algoritmo%20Gradient%20Boosting%20é%20uma%20técnica%20de%20aprendizado%20de,fracos%2C%20geralmente%20árvores%20de%20decisão. Acesso em: set. 2022.

WANG, Jian. BERZINS, Kaspars. HICKS, Diana. MELKERS, Júlia. XIAO, Fang. PINHEIRO, Diogo**. A boosted-trees method for name disambiguation.** Disponível em: https://link.springer.com/article/10.1007/s11192-012-0681-1. Acesso em: mar. 2022.

WEIS, M. NAUMANN, F**.** Duplicate detection in XML. **In: International Workshop In Information Quality In Information Systems,** 1.,2004, Paris, França. Proceedings...New York, NY: ACM. 2004, p. 10-191

WEIS, m.; NAUMANN, F. Relationship Based Duplicate Detection. **Technical Report** Nr. Hu-1B-206.[S.1.:s.n], 2006.

ZHU, Jia. Wu Xingcheng. LIN, Xueqin. HUANG, Changgin. PUI, Gabriel. FUNG, Cheong. TANG, Young. **A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering.** Disponível em: https://link.springer.com/article/10.1007/s11192-017-2611-8. Acesso em: mar. 2022.