25°enancib
Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
Informação, Decolonialidade e Direitos Difusos para o Desenvolvimento Sustentável
3 a 7 de novembro de 2025 - Rio de Janeiro

XXV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - XXV ENANCIB

GT 7 – Produção e Comunicação da Informação em Ciência, Tecnologia & Inovação

RESOLUÇÃO DE ENTIDADES PARA POLÍTICAS DE CT&I: UM MÉTODO BASEADO EM IA PARA CRUZAMENTO DE DADOS ADMINISTRATIVOS E CURRICULARES

ENTITY RESOLUTION FOR S&T&I POLICIES: *AN AI-BASED METHOD FOR MATCHING ADMINISTRATIVE AND CURRICULUM DATA*

Iago Breno Alves do Carmo Araujo - Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)
Jesús P. Mena-Chalco - Universidade Federal do ABC (UFABC)

Modalidade: Trabalho Completo

Resumo: A qualidade dos dados institucionais é importante para políticas eficazes em Ciência, Tecnologia e Inovação (CT&I). Na Ciência da Informação, garantir dados confiáveis é essencial para estudos bibliométricos e avaliação científica. Este trabalho propõe uma metodologia para resolução de entidades, que verifica se pesquisadores possuem afiliações institucionais coerentes com registros administrativos. A abordagem cruza informações de bases administrativas de financiamento à pesquisa com currículos acadêmicos, considerando o histórico de afiliações profissionais. Busca-se correspondência em nível institucional e, quando possível, em unidades acadêmicas específicas, levando em conta o período ativo dos processos. O método integra técnicas tradicionais de similaridade textual com abordagens baseadas em Inteligência Artificial, especialmente Modelos de Linguagem de Grande Escala, aplicados em aprendizado com poucos exemplos (*few-shot learning*) e complementados por buscas automatizadas na web. Os resultados indicam eficácia significativa frente a dados incompletos ou inconsistentes, apresentando uma solução robusta e escalável para validação cruzada em ambientes heterogêneos. A combinação entre métodos tradicionais e tecnologias de IA proporciona uma avaliação abrangente da concordância entre registros administrativos e curriculares, aprimorando práticas de gestão e curadoria informacional em CT&I.

Palavras-chave: Resolução de entidades; Dados institucionais; Modelos de Linguagem; Automatização

Abstract: The quality of institutional data is crucial for the development of effective Science, Technology, and Innovation (ST&I) policies. In Information Science, ensuring reliable data is essential for bibliometric studies and scientific assessment. This paper proposes a methodology for entity resolution that verifies whether researchers have institutional affiliations consistent with administrative records. The approach cross-references information from administrative research funding databases with academic curricula, taking into account the history of professional affiliations. It seeks to match entities at the institutional level and, when possible, at the level of specific academic units, considering the active period of each record. The method integrates traditional text similarity techniques with Artificial Intelligence approaches, particularly Large Language Models, applied in few-shot learning settings and complemented by automated web searches. The results indicate significant effectiveness when handling incomplete or inconsistent data, providing a robust and scalable solution for cross-validation in heterogeneous environments. The combination of traditional methods with AI technologies enables a comprehensive assessment of the concordance between administrative and curriculum records, enhancing management and information curation practices in ST&I.

Keywords: Entity resolution; Institutional records; Language models; Automation.

1 INTRODUÇÃO

A resolução de entidades consiste na tarefa de identificar diferentes registros presentes em uma ou mais bases de dados que se referem à mesma entidade do mundo real. Esta tarefa é um desafio frequentemente abordado no contexto da organização da informação, pois diferentes fontes podem registrar a mesma entidade sob formas variadas e inconsistentes (Christen, 2012). Este processo é importante para assegurar a qualidade das informações e apoiar decisões estratégicas baseadas em dados confiáveis, especialmente em contextos que demandam precisão, como bibliotecas digitais e plataformas acadêmicas.

Na literatura científica, a resolução de entidades é conhecida por diferentes denominações, entre elas destacam-se: deduplicação (*deduplication*), ligação de registros (*record linkage*), correspondência de entidades (*entity matching*) e desambiguação de nomes (*name disambiguation*) (Christophides *et al.*, 2021; Peeters; Bizer, 2023). Embora os termos possam variar conforme a área de estudo ou aplicação específica, todos remetem essencialmente ao mesmo desafio conceitual: assegurar que registros heterogêneos sejam associados corretamente à entidade que representam.

No contexto específico da Ciência da Informação, problemas clássicos frequentemente encontrados estão relacionados à ambiguidade e variações na representação dos nomes de autores, bem como informações institucionais incompletas ou inconsistentes, ou títulos de publicações diferentes (Mena-Chalco *et al.*, 2024). Autores com nomes muito comuns (homônimos), diferentes variações de um mesmo nome (heterônimos), além de afiliações institucionais desatualizadas, são fatores que dificultam a correta identificação e agrupamento das informações acadêmicas e bibliométricas (Shin *et al.*, 2014; Binette *et al.*, 2024). Esses fatores afetam significativamente a qualidade e confiabilidade das análises bibliométricas, potencialmente impactando decisões estratégicas em ciência, tecnologia e inovação (CT&I).

Embora o problema de resolução de entidades não seja recente e diversos métodos tenham sido propostos, ainda persistem desafios devido à crescente complexidade e ao volume de dados presentes nos dias de hoje. Tradicionalmente, as abordagens incluem desde métodos baseados em regras até técnicas estatísticas avançadas, passando pela combinação de diferentes atributos como, por exemplo, coautoria, título e informações institucionais

(Mudgal *et al.*, 2018). Mais recentemente, técnicas baseadas em inteligência artificial têm ganhado destaque, especialmente o uso de modelos de linguagem generativos, como o GPT, que oferecem maior flexibilidade e eficácia na resolução desses problemas, mesmo com poucos dados de treinamento (Peeters; Bizer, 2023).

Neste contexto, este trabalho propõe uma metodologia, com base computacional, escalável e robusta de resolução de entidades, especificamente voltada à verificação da coerência entre as **afiliações institucionais de pesquisadores** registradas em bases administrativas e currículos acadêmicos. Este problema de afiliações é comum quando se tenta relacionar informação presente nas instituições acadêmicas com a informação presente nos currículos dos consultados, (e.g. instituição do projeto de pesquisa registrado na instituição acadêmica *versus* a instituição do projeto de pesquisa registrado no CV Lattes).

A abordagem apresentada combina métodos tradicionais de similaridade textual com abordagens modernas baseadas em modelos de linguagem de grande escala (*Large Language Models*, LLMs) (Raaian *et al.*, 2024). Utilizando técnicas de aprendizado com poucos exemplos (*few-shot learning*) (Wang *et al.*, 2021) e buscas automatizadas na web, busca-se garantir uma correspondência precisa tanto em nível institucional quanto em unidades acadêmicas específicas, considerando também o período ativo dos processos.

A importância de soluções escaláveis para a resolução de entidades fica evidente frente à quantidade massiva de dados produzidos e acumulados nas instituições acadêmicas (e.g., Universidade Federal) e de fomento à pesquisa (e.g., Fundação de Amparo e Auxílio Financeiro - FAPs).

A metodologia proposta neste trabalho é validado através de um estudo de caso aplicado a bases administrativas de financiamento à pesquisa e currículos acadêmicos, demonstrando resultados significativamente positivos mesmo em cenários com dados incompletos ou inconsistentes. Estes resultados preliminares apontam para uma melhoria expressiva na qualidade e confiabilidade dos dados administrativos e curriculares.

O presente estudo contribui potencialmente para o desenvolvimento teórico e metodológico na área da Ciência da Informação e Bibliometria/Cientometria, assim como para o aperfeiçoamento das ferramentas e processos que dão suporte às políticas de gestão científica (como, por exemplo, as políticas de gestão dentro das FAPs).

Diante desse cenário, o presente estudo busca responder ao seguinte problema de pesquisa: como verificar a coerência entre as afiliações institucionais de pesquisadores

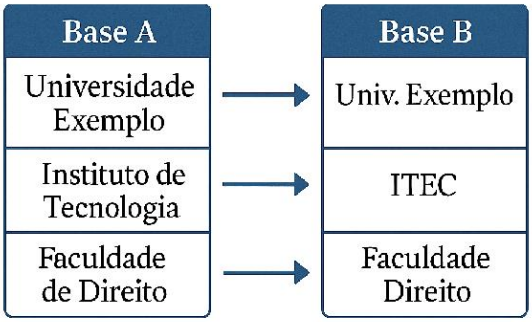
registradas em bases administrativas de financiamento e aquelas declaradas em currículos acadêmicos? Para enfrentar esse desafio, propõe-se um método híbrido de resolução de entidades que combina técnicas de similaridade textual com modelos de linguagem de grande escala, avaliando sua efetividade em um estudo de caso aplicado a dados reais. A principal contribuição do trabalho é oferecer uma abordagem para aprimorar a qualidade e a confiabilidade dos dados utilizados em análises bibliométricas e de políticas de CT&I. Além desta introdução, o artigo está estruturado da seguinte forma: a seção 2 discute o problema de cruzamento de dados; a seção 3 apresenta o método proposto; a seção 4 descreve e analisa o estudo de caso; e a seção 5 traz as considerações finais.

2 ENTENDENDO O PROBLEMA DE CRUZAMENTO DE DADOS

A resolução de entidades é uma tarefa relevante quando lidamos com dados oriundos de fontes distintas. No contexto da identificação de instituições, o problema é particularmente complexo devido à variedade de formas como os nomes das instituições são registrados. No **primeiro nível**, o objetivo é identificar se nomes distintos como “Universidade Exemplo” e “Univ. Exemplo” se referem à mesma universidade. Já no **segundo nível,** o objetivo é verificar se unidades como “Instituto de Tecnologia” e “ITEC” representam o mesmo instituto ou faculdade pertencente àquela universidade (ou seja do primeiro nível).

As fontes A e B podem vir de sistemas distintos, com padrões heterogêneos de nomenclatura, abreviações, erros ortográficos ou estruturas diferentes de representação hierárquica (ver Figura 1). Isso exige técnicas que levem em conta tanto a semelhança lexical quanto o contexto institucional. Muitas vezes, a tarefa requer uma análise em dois estágios. O primeiro consiste em identificar a universidade principal, enquanto o segundo envolve verificar se as unidades subordinadas são equivalentes. Isso se dá porque muitas unidades, no segundo nível, podem ter o mesmo nome.

Figura 1 - Exemplo de cruzamento de dados entre duas bases diferentes (Bases A e B). A seta indica a correspondência de A em B



Fonte: Elaborado pelos autores (2025).

O desafio consiste, portanto, em construir um sistema capaz de inferir que dois conjuntos de nomes se referem a uma mesma entidade real, mesmo quando escritos de maneira distinta. Essa tarefa é central em projetos de integração de dados acadêmicos, bases curriculares, repositórios de pesquisa e sistemas governamentais, onde a precisão na correspondência institucional impacta diretamente a qualidade das análises e das decisões. Na próxima seção, discute-se o método proposto para lidar com todos os aspectos envolvidos na identificação de instituições tanto no primeiro nível quanto no segundo nível.

3 NOVO MÉTODO DE RESOLUÇÃO DE ENTIDADES BASEADO EM IA

A resolução de entidades consiste na tarefa de identificar diferentes registros que, embora possam parecer distintos à primeira vista, referem-se, na verdade à mesma entidade real. Isso é especialmente relevante para a Ciência da Informação, dado que informações corretas e bem estruturadas são fundamentais para estudos bibliométricos e análises institucionais. Este artigo propõe um método computacional para resolver entidades utilizando técnicas híbridas que combinam estratégias tradicionais de comparação textual com o poder analítico de modelos recentes de Inteligência Artificial.

O novo método proposto compreende três etapas principais, descritas detalhadamente a seguir: (1) Normalização dos dados; (2) Identificação da similaridade textual; e (3) Cruzamento por meio de Modelos de Linguagem de Grande Escala.

3.1 Normalização dos dados

A primeira etapa tem como objetivo preparar os dados para análise, reduzindo variações ortográficas e diferenças na formatação que possam prejudicar a identificação

correta das entidades. Inicialmente, os textos dos registros são convertidos para letras minúsculas e espaços extras são eliminados. Em seguida, caracteres especiais, como letras acentuadas, são substituídos por suas versões básicas (por exemplo, "á" se torna "a"), visando maior uniformidade. Adicionalmente, barras e sinais de pontuação foram substituídos ou removidos para simplificar o processo subsequente de análise textual. Ao final dessa etapa, obteve-se um conjunto de dados uniformizado que facilita a comparação entre registros.

3.2 Identificação da similaridade textual

Após a normalização dos dados, aplicou-se uma técnica tradicional chamada TF-IDF (Aizawa, 2003), amplamente utilizada em Ciência da Informação para comparação textual. Essa técnica transforma textos em vetores numéricos, permitindo medir o quanto duas entidades são semelhantes com base em seus termos e estruturas internas.

Nesta fase, o método tratou valores faltantes de forma consistente, convertendo-os em cadeias vazias. Para otimizar o desempenho computacional, apenas valores únicos foram analisados. A técnica de vetorização foi realizada considerando pequenos grupos de caracteres, n-gramas de três a cinco caracteres, de modo a capturar padrões importantes, mesmo que pequenas variações textuais existissem entre os registros. A similaridade entre esses vetores foi calculada utilizando a distância ”cosseno” (Usino *et al.*, 2019), uma medida matemática simples e eficiente que indica o grau de proximidade entre dois textos. Um limiar de similaridade de 0,85 foi estabelecido, classificando automaticamente como correspondentes diretos (cruzamentos fortes) aqueles pares de entidades que superassem esse valor. Registros com similaridade menor são encaminhados para uma análise mais detalhada.

3.3 Cruzamento por meio de Modelos de Linguagem de Grande Escala

Para lidar com casos em que a similaridade textual tradicional não for suficiente, considera-se um modelo avançado de IA denominado modelo de linguagem (e.g., Gemini-2.0-Flash), que também faz consultas adicionais na web para obter contexto adicional sobre os dados.

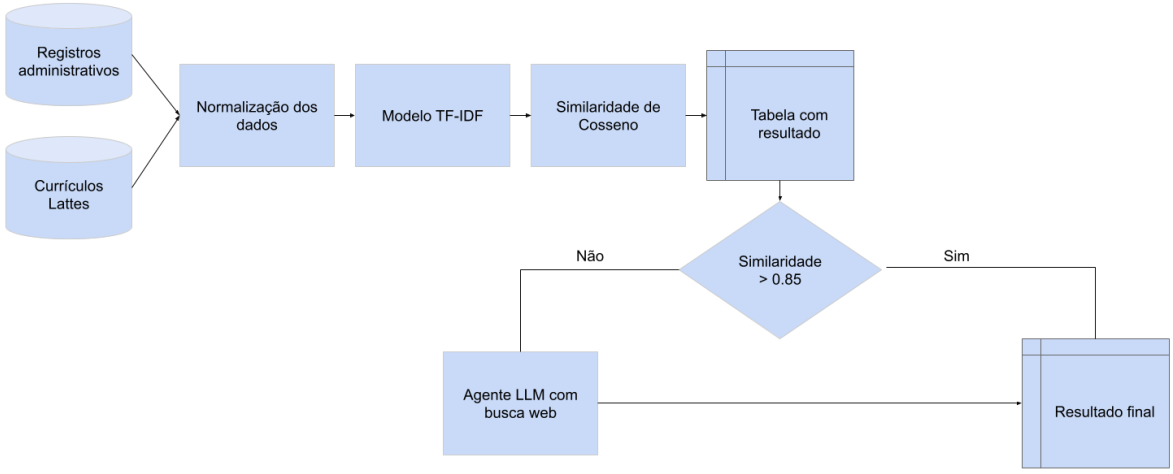
O processo envolve apresentar ao modelo o nome da entidade (por exemplo, uma instituição ou unidade acadêmica) e detalhes adicionais relevantes, solicitando que identificasse a correspondência mais provável entre diferentes registros. Para cada consulta,

o modelo de IA retorna uma resposta estruturada contendo o nome da unidade identificada, uma indicação clara (sim ou não) sobre a correspondência e uma breve explicação justificando a decisão tomada.

Os resultados da análise por TF-IDF foram então combinados com as respostas geradas pelo modelo de IA. Um limiar final de 0,85 foi utilizado para aceitação direta dos resultados. Os casos com valores inferiores a este limiar foram encaminhados ao julgamento semântico e contextual do modelo de IA. Essa estratégia híbrida assegura uma primeira triagem eficiente, seguida por uma análise mais aprofundada para casos mais difíceis.

Este método híbrido mostra-se bastante promissor em contextos de Ciência da Informação, sobretudo em bases de dados complexas e heterogêneas, onde diferenças semânticas podem não ser captadas plenamente por métodos tradicionais de comparação textual. Ao integrar técnicas consolidadas como o TF-IDF com as potencialidades das novas gerações de IA, o modelo proposto permitiria uma resolução de entidades mais precisa e aplicável em larga escala, preservando a qualidade das análises e estudos bibliométricos.

Figura 2 – Procedimentos considerados no novo método de cruzamento de dados (método híbrido)



Fonte: Elaborado pelos autores (2025).

Finalmente, a Figura 2 apresenta visualmente todos os procedimentos do método, destacando como as estratégias tradicionais são complementadas pelas técnicas baseadas em IA para resolver de forma robusta a resolução de entidades.

4 ESTUDO DE CASO

O procedimento de verificação da concordância seguiu uma lógica em etapas. Para cada registro de auxílio ou bolsa presente na base administrativa, identificou-se o pesquisador responsável (ou contemplado) e a instituição de vínculo registrada nesse processo. Em seguida, buscou-se o currículo Lattes correspondente a esse pesquisador, extraindo-se a instituição de vínculo declarada no mesmo período. A comparação foi então realizada entre a instituição de vínculo registrada no auxílio e a instituição informada no currículo. Quando disponível, também foi considerada a correspondência em nível hierárquico interno (unidades acadêmicas), de modo a oferecer maior granularidade na análise.

Para avaliar a efetividade do novo método de resolução de entidades aqui proposto, foi conduzido um estudo de caso com dados reais provenientes de uma agência pública de fomento à pesquisa. Foram utilizados registros administrativos de concessão de auxílios e bolsas, combinados com informações extraídas de currículos disponíveis na Plataforma Lattes, mantida pelo CNPq. A base analisada inclui 77.069 registros de auxílios a projetos de pesquisa e 13.826 registros de bolsas (ao todo 90.895 registros). Estavam associados a esses registros cerca de 31 mil pesquisadores. O objetivo foi verificar o grau de concordância entre as instituições informadas nos registros administrativos e aquelas registradas nos currículos Lattes, utilizando o método híbrido desenvolvido, que combina comparação textual por TF-IDF e análise semântica com modelos de linguagem.

No **primeiro nível** de análise dos auxílios (instituição-sede), o componente baseado em TF-IDF identificou cruzamento direto em 80% dos casos (61.820 de 77.069 registros). Para os 20% restantes, o modelo de linguagem foi acionado, obtendo sucesso em **19,48%** dos casos. Assim, o método atingiu 99,48% de concordância total, com apenas 0,52% (390 registros) não correspondidos, casos em que o modelo indicou, com justificativa, a ausência de correspondência real. Na análise do **segundo nível** (considerando as unidades dentro das instituições), o TF-IDF obteve sucesso em **72,8%** dos registros (49.744 de 68.415). O modelo de linguagem contribuiu com 18,6%, totalizando 91,4% de cruzamento identificado. Permaneceram sem correspondência 8,6% dos casos (5.898 registros), geralmente por falta de informação nas bases.

Aqui é importante frisar que, mesmo quando não havia informação sobre a unidade, o desempenho do método se manteve elevado: 95% de cruzamento identificado (8.429 de

8.871), sugerindo que a ausência de detalhamento não comprometeu a capacidade de inferência do modelo. No caso dos registros de bolsas, o método também demonstrou robustez, embora com variações. No nível institucional, o TF-IDF identificou 62,3% dos cruzamentos (8.616 de 13.826), enquanto o modelo de linguagem adicionou mais 29,9% (4.132), alcançando 92,2% de concordância. No nível que considera as unidades, o TF-IDF obteve 55,6% (7.023 de 12.623), e o modelo contribuiu com mais 26,2% (3.298), totalizando 73,8%. Para registros de bolsas sem informação de unidade, a divisão dos acertos foi equilibrada: 52,7% resolvidos por LLMs e 47,3% por TF-IDF.

Esses resultados indicam que o método híbrido é eficaz em diferentes níveis de estrutura dos dados. O modelo de linguagem Gemini-2.0-Flash mostrou especial capacidade de lidar com registros incompletos ou inconsistentes, fornecendo inferência contextual onde os métodos tradicionais falham. A comparação entre os conjuntos de auxílios e bolsas também revela que bases com maior padronização favorecem abordagens lexicais, enquanto bases mais heterogêneas se beneficiam do uso de modelos de linguagem.

As Tabelas 1 e 2 detalham os percentuais de concordância obtidos entre os registros administrativos e os dados oriundos dos currículos Lattes, considerando diferentes níveis de granularidade institucional (entidade, unidade e ausência de unidade) para os casos de auxílios e bolsas, respectivamente.

Tabela 1 – Resultado da concordância entre os registros administrativos de auxílios e CVs dos pesquisadores

Caso / Método	**TF-IDF**	**Agente LLM**	**Total**
Auxílio Entidade	80% (61.820) / 20% (15.466)	19.48% (14.859) / 0.52% (390)	99.48% (76.679) / 0.52% (390)
Auxílio Unidade	72.8% (49.744) / 27.2% (18.641)	18.6% (12.743) / 8.6% (5.898)	91.4% (62.517) / 8.6% (5.898)
Auxílio Sem Unidade	-	95.0% (8.429) / 5.0% (442)	95.0% (8.429) / 5.0% (442)

Fonte: Elaborado pelos autores (2025).

Nessas Tabelas os valores estão destacados em cores para facilitar a leitura e compreensão dos resultados. A cor azul representa os percentuais e totais de registros que foram corretamente cruzados (isto é, aqueles que encontraram correspondência entre os registros administrativos e os dados dos currículos) em cada etapa do processo. Já a cor vermelha indica os registros que não foram resolvidos em determinada etapa, ou seja, os casos que ainda permaneceram sem correspondência após a aplicação do respectivo método.

Tabela 2 – Resultado da concordância entre os registros administrativos de bolsas e CVs dos pesquisadores

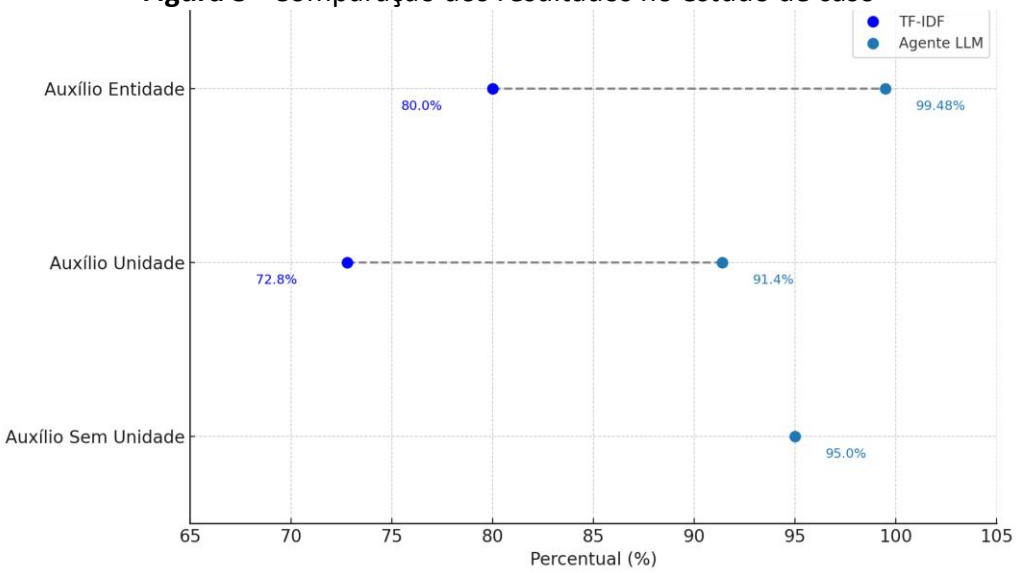
Caso / Método	TF-IDF	Agente LLM	Total
Bolsa Entidade	62.3% (8.616) / 37.7% (5.210)	7.8% (1.078) / 29.9% (4.132)	70.1% (9.694) / 29.9% (4.132)
Bolsa Unidade	55.6% (7023) / 44.4% (5.600)	18.2% (2.302) / 26.2% (3.298)	73.8% (9.325) / 26.2% (3.298)
Bolsa Sem Unidade	-	52.7% (634) / 47.3% (569)	52.7% (634) / 47.3% (569)

Fonte: Elaborado pelos autores (2025).

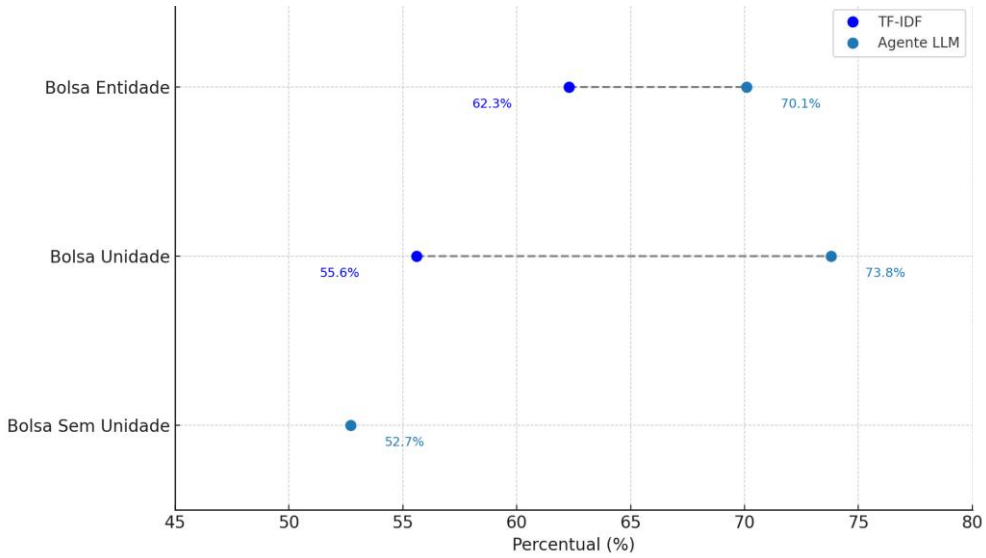
Finalmente, a Figura 3(a) apresenta graficamente os resultados obtidos para os **registros de auxílios**, comparando os percentuais de cruzamento alcançados por meio do TF-IDF (representado pelos pontos em azul mais escuro, à esquerda) e o total obtido após a complementação com o agente baseado em modelo de linguagem (em azul claro, à direita). As linhas tracejadas indicam a diferença entre o desempenho inicial do método tradicional e o ganho proporcionado pela etapa semântica. Observa-se que, tanto no nível de entidade quanto de unidade, há um aumento expressivo na taxa de correspondência após o uso do agente LLM, passando, por exemplo, de 80% para 99,48% no nível de entidade, e de 72,8% para 91,4% no nível de unidade. No caso de registros sem unidade, apenas o agente LLM atuou, alcançando 95% de sucesso.

Já a Figura 3(b) apresenta graficamente os resultados obtidos para os **registros de bolsas**. Também nesse cenário observa-se o ganho incremental promovido pela aplicação do modelo de linguagem. No nível de entidade, o percentual sobe de 62,3% (TF-IDF) para 70,1% no total. No nível de unidade, o avanço é ainda mais expressivo, de 55,6% para 73,8%. Por fim, no nível sem unidade, o resultado final (52,7%) reflete exclusivamente a atuação do agente LLM, indicando sua relevância em contextos com dados mais limitados.

Figura 3 - Comparação dos resultados no estudo de caso



(a) Registros sobre auxílios



(b) Registros sobre entidades.
Fonte: Elaborado pelos autores.

A diferença de desempenho entre os conjuntos de auxílios e de bolsas pode ser explicada pelo grau de padronização presente em cada base. Os registros de auxílios, por envolverem processos institucionais mais formais e estruturados, tendem a apresentar maior consistência na indicação da instituição-sede. Já os registros de bolsas mostram maior heterogeneidade, seja pela diversidade de tipos de vínculo (iniciação científica, mestrado, doutorado, pós-doutorado), seja pela forma como os dados foram historicamente coletados e registrados. Nesses casos, em que há mais variação e inconsistência textual, os modelos de linguagem demonstraram maior efetividade ao complementar as abordagens lexicais tradicionais.

Apesar dos resultados relevantes, algumas limitações foram observadas. Parte dos registros permaneceu sem correspondência, em especial aqueles com informações ausentes ou desatualizadas nos currículos Lattes. Também se verificaram dificuldades decorrentes da diferença de granularidade entre as bases: enquanto os registros administrativos frequentemente indicam apenas a instituição-sede, os currículos nem sempre detalham a unidade vinculada, o que compromete análises em nível hierárquico mais refinado. Além disso, a utilização de LLMs em larga escala pode implicar custos computacionais consideráveis, o que pode restringir a aplicação do método em bases muito volumosas.

5 CONSIDERAÇÕES FINAIS

Um dos principais desafios enfrentados na gestão e análise de dados em Ciência da Informação ocorre quando há necessidade de integrar informações provenientes de fontes distintas. A ausência de correspondência direta entre registros (como, por exemplo, dados administrativos e currículos acadêmicos) compromete a confiabilidade das análises, limita a automação de processos e dificulta a construção de indicadores robustos. Esse problema, conhecido como resolução de entidades, não é novo, mas se intensifica à medida que aumenta a complexidade e a heterogeneidade das bases informacionais utilizadas.

Tradicionalmente, técnicas baseadas em comparação lexical (como a vetorização por TF-IDF) têm sido amplamente empregadas para resolver esse tipo de inconsistência. Contudo, em muitos contextos, especialmente quando os dados apresentam variações semânticas ou são incompletos, tais abordagens mostram-se insuficientes. Neste cenário, os avanços recentes em IA abrem novas possibilidades. Ao incorporar capacidade de inferência contextual e compreensão semântica, esses modelos oferecem uma camada complementar que potencializa a precisão do cruzamento entre registros distintos (ainda mais usando consultas automatizadas na web).

O estudo de caso apresentado neste trabalho evidenciou, de forma empírica, a eficácia de um método híbrido que alia estratégias clássicas e modelos de linguagem para resolver inconsistências entre registros de auxílios e bolsas e os dados da Plataforma Lattes. Os resultados indicam altos índices de concordância, mesmo em cenários com ausência de informação detalhada, o que reforça a aplicabilidade prática do método.

Como trabalho futuro, propõe-se aprofundar a investigação sobre o modelo de decisão adotado, incorporando princípios da Inteligência Artificial Explicável (Marcus; Teuwen, 2024).

Compreender de forma transparente como os modelos de linguagem realizam suas inferências é relevante para garantir confiança, rastreabilidade e validação das decisões automatizadas, especialmente em contextos institucionais da Ciência da Informação.

É importante destacar que, embora o estudo aqui tenha se concentrado na correspondência entre instituições, o método proposto pode ser adaptado para outros tipos de entidades, como nomes de pesquisadores, grupos de pesquisa, áreas de atuação, ou ainda para dados oriundos de outros domínios informacionais. Trata-se, portanto, de uma contribuição em termos de método com potencial de reaplicação em cenários diferentes.

Finalmente, apesar de iniciativas internacionais, como o Research Organization Registry (RoR), avancem no sentido de oferecer identificadores persistentes para instituições de pesquisa, sua adoção ainda enfrenta limitações práticas. Em muitos casos, a atualização não contempla diferentes níveis hierárquicos, como a distinção entre instituição-sede e suas unidades acadêmicas internas. Essa lacuna compromete a padronização necessária para análises mais refinadas, reforçando a importância de métodos complementares de resolução de entidades que integrem identificadores globais a técnicas de normalização e inferência contextual.

REFERÊNCIAS

AIZAWA, Akiko. An information-theoretic perspective of tf–idf measures. **Information Processing & Management**, v. 39, n. 1, p. 45-65, 2003.

BINETTE, Olivier; *et al*. How to evaluate entity resolution systems: an entity-centric framework with application to inventor name disambiguation. **Arxiv.org,** 2024. Disponível em: https://arxiv.org/abs/2404.05622v1. Acesso em: 23 maio 2025.

CHRISTEN, Peter. **Data matching**: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin: Springer, 2012.

CHRISTOPHIDES, Vassilis; *et al*. An overview of end-to-end entity resolution for big data. **ACM Computing Surveys** (CSUR), v. 53, n. 6, 2021.

MARCUS, Eric; TEUWEN, Jonas. Artificial intelligence and explanation: how, why, and when to explain black boxes. **European Journal of Radiology**, v. 173, 111393, 2024. ISSN 0720-048X. Disponível em: https://doi.org/10.1016/j.ejrad.2024.111393. Acesso em: 20 maio 2025.

MENA-CHALCO, Jesus P.; *et al*. O desafio da deduplicação de publicações: criação e avaliação de um benchmark. *In*: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, Brasília, 2024. **Anais** [...]. Brasília: UnB/Ibict, 2024. Disponível em:

https://ridi.ibict.br/bitstream/123456789/1323/1/O%20desafio%20da%20deduplica%C3%A7%C3%A3o%20de%20publica%C3%A7%C3%B5es.pdf. Acesso em: 26 jan. 2026.

MUDGAL, Sidharth; *et al*. Deep learning for entity matching: a design space exploration. *In*: SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2018, Houston. **Proceedings** […]. New York: ACM, 2018. p. 19-34. Disponível em: https://pages.cs.wisc.edu/~anhai/papers1/deepmatcher-sigmod18.pdf. Acesso em: 26 jan. 2026.

RAAIAN, Mohaimenul Azam Khan; *et al*. A review on large language models: architectures, applications, taxonomies, open issues and challenges. **IEEE Access**, v. 12, p. 26839–26874, 2024. Disponível em: https://ieeexplore.ieee.org/document/10433480. Acesso em: 26 jan. 2026.

PEETERS, Ralph; BIZER, Christian. Using ChatGPT for entity matching. **Arxiv.org**, 2023. Disponível em: https://arxiv.org/abs/2305.03423v2. Acesso em: 23 maio 2025.

SHIN, Dongwook; *et al*. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. **Scientometrics**, v. 100, n. 1, p. 15-50, 2014. Disponível em: https://link.springer.com/article/10.1007/s11192-014-1289-4. Acesso em: 26 jan. 2026.

USINO, Wendi; *et al*. Document similarity detection using k-means and cosine distance. **International Journal of Advanced Computer Science and Applications**, v. 10, n. 2, 2019.

WANG, Yaking; *et al*. Generalizing from a few examples: a survey on few-shot learning. **ACM Computing Surveys**, v. 53, n. 3, art. 63, p. 1–34, maio 2021. Disponível em: https://doi.org/10.1145/3386252. Acesso em: 20 maio 2025.