UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA

ARMANDO BARBOSA SOBRINHO

ALINHAMENTO DE DADOS CONECTADOS A PARTIR DE CONCEITOS PRIMÁRIOS E SECUNDÁRIOS

Maceió-AL

2016

ARMANDO BARBOSA SOBRINHO

ALINHAMENTO DE DADOS CONECTADOS A PARTIR DE CONCEITOS PRIMÁRIOS E SECUNDÁRIOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Ig Ibert Bittencourt Santana Pinto
Coorientador: Sean Wolfgand Matsui Siqueira

Maceió-AL

2016

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário Responsável: Valter dos Santos Andrade

B238a Barbosa Sobrinho, Armando.
Alinhamento de dados conectados a partir conceitos primários e secundários / Armando Barbosa Sobrinho. – 2017.
62 f.: il.

Orientador: Ig Ibert Bittencourt Santana Pinto.
Coorientador: Sean Wolfgand Matsui Siqueira.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Programa de Pós-Graduação em Informática. Maceió, 2017.

Bibliografia: f. 60-62.

1. Correspondência de instância. 2. Alinhamento de dados. 3. Datasets. 4. Dados conectados. 5. Web de dados. I. Título.

CDU: 004.738.5.057

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Membros da Comissão Julgadora da Dissertação de Mestrado de Armando Barbosa Sobrinho, intitulada: *"Alinhamento de Dados Conectados a partir de Conceitos Primários e Secundários"*, apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas em 01 de fevereiro de 2017, às 14h00min, no Miniauditório do CEPETEC no Instituto de Computação da UFAL.

COMISSÃO JULGADORA

Prof. Dr. Ig Ibert Bittencourt Santana Pinto
UFAL – Instituto de Computação
Orientador

Prof. Dr. Sean Wolfgand Matsui Siqueira
UNIRIO – Universidade Federal do Estado do Rio de Janeiro
Orientador

Prof. Dr. Rafael de Amorim Silva
UFAL – Instituto de Computação
Examinador

Prof. Dr. Bernardo Pereira Nunes
PUC-Rio – Pontifícia Universidade Católica do Rio de Janeiro
Examinador

À Deus, pois sem sua força não conseguiria superar meus obstáculos.
Aos meus pais Joana e Ednaldo,
por sempre estarem comigo em todos os momentos.

AGRADECIMENTOS

Agradeço à Deus.

Agradeço aos meus pais, minha namorada e família por me apoiarem moral e financeiramente para que este dia se tornasse realidade.

Agradeço ao meus orientadores, Ig e Sean, por me ensinarem o significado de resiliência, por todos os conselhos, pela paciência e ajuda nesse período.

Aos meus amigos, Álvaro, Adolfo, Carlos, Eduarddo, Guilherme, Manoel, Samário. Em especial ao André, Josmário, Judson e Williams, por abrirem mão de suas responsabilidades e ajudarem do desenvolvimento esta dissertação.

Ao NEES, por mostrar que o conhecimento deve ser compartilhado e transformado em ações com impacto social.

A todos os professores, pois através deles pude ver o mundo.

À FAPEAL pelo apoio financeiro para realização deste trabalho de pesquisa.

“Eu acredito demais na sorte.
E tenho constatado que,
quanto mais duro eu trabalho,
mais sorte eu tenho.”
(Thomas Jefferson)

RESUMO

Nos últimos anos, dados conectados têm sido a forma mais proeminente para abertura de dados em diversos países. Tal forma utiliza padrões para descrição destes dados, promovendo a sua interoperabilidade, seu reuso e a sua integração. No entanto, integrar a informação entre diferentes conjuntos de dados surge como um empecilho para o seu desenvolvimento, principalmente se tal integração consistir na correspondência de uma determinada entidade do mundo real em conjuntos de dados distintos. Neste contexto, este trabalho propõe uma abordagem para auxiliar na identificação de instâncias correspondentes. Para isso, baseia-se na modelagem conceitual dos dados, permitindo que os relacionamentos entre os conceitos sejam utilizados para descobrir novas correspondências entre os dados. Para avaliar a eficácia da proposta foram realizados um estudo de caso e um experimento. No estudo de caso, a proposta foi utilizada para encontrar as correspondências de pesquisadores e publicações em quatro *datasets* (Lattes, RBIE, SBIE e WIE) e, então, responder um conjunto com mais de trinta perguntas realizadas pela comunidade de Informática na Educação. No experimento, a proposta foi utilizada em dois cenários (C1 e C2) e comparada a outras abordagens através das métricas de precisão, revocação e medida-f. De acordo com os resultados apresentados, a proposta posicionou-se em primeiro e segundo lugar nos cenários C1 e C2 respectivamente, mesmo não utilizando computações específicas para os *datasets*, permitindo sua utilização em outros contextos com o mínimo de esforço.

Palavras-chaves: Correspondência de Instâncias, Alinhamento de Dados, *Datasets*, Dados Conectados, Web de Dados.

ABSTRACT

Recently linked data has been the most prominent way to open data in several countries. This way it uses standards to describe this data, promoting its interoperability, its reuse and its integration. However, integrating information between different datasets is a hindrance to their development, especially if such integration consists of matching a particular real-world entity in distinct datasets. In this context, this work proposes an approach to assist in the identification of corresponding instances. For this, it is based on the conceptual modeling of the data, allowing the relationships between the concepts to be used to discover new correspondences between the data. To evaluate the effectiveness of the proposal, a case study and an experiment were carried out. In the case study, the proposal was used to find the correspondence of researchers and publications in four datasets (Lattes, RBIE, SBIE and WIE). These correspondences was used to answer a set with more than thirty questions provided by the informatics community in Education. In the experiment, the proposal was used in two scenarios (C1 and C2) and compared to other approaches through precision, recall and f-measure metrics. According to the presented results, the proposal ranked first and second place in scenarios C1 and C2 respectively, even though it did not use specific computations for the datasets, allowing its use in other contexts with the least effort.

KEY-WORDS: Instance Matching, Data Correspondence, Dataset, Linked Data, Web of Data

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A Web é baseada em documentos, onde os documentos ou páginas são descritas através de uma linguagem de marcação de hipertexto (HTML) e interligados através de *hyperlinks*. Para acessar as páginas na Web, é necessário um mecanismo capaz de identificar de forma única cada uma dessas páginas, o que é feito através de um mecanismo de identificação global chamado URI (*Uniform Resource Identifier*). Dessa forma, para ter acesso a um conteúdo específico de qualquer documento neste paradigma, é necessário obter todo o documento.

Analogamente, a Web Semântica, ou Web de dados[1], através de URIs, possibilita que recursos na web sejam identificados de forma única, viabilizando assim a recuperação de seus conteúdos. Esses, por sua vez, são representados através de um framework de descrição de recursos (RDF), além de contar com um protocolo para consulta, o *Simple Protocol and RDF Query Language* (SPARQL). Tais elementos, juntamente com princípios e boas práticas, surge o conceito de Dados Conectados.

Dados Conectados estão intimamente relacionados à Web de dados, pois de acordo com Bizer, Heath e Berners-Lee (2009), dados de diferentes fontes são disponibilizados na Web e conectados uns aos outros. Assim, Dados Conectados são considerados um ponto chave para o desenvolvimento da Web Semântica (BERNERS-LEE et al., 2001), Além de seu potencial uso para negócios e governos (ISOTANI; BITTENCOURT, 2015).

Na perspectiva de negócio, temos o caso da *Best Buy* (BRINKER, 2009), que através da utilização de serialização RDF[2] (RDFa[3]) melhorou o número de acessos via buscador Google entre 15% e 30%. E do Google, que passou a utilizar a serialização json-ld[4] em um de seus produtos, o Gmail (SIMISTER; BRICKLEY, 2013). Com isso, a aplicação passou a categorizar e fornecer visualizações personalizadas de acordo com o e-mail recebido. No governo, temos o caso do Reino Unido que publica seus dados em formato RDF (SHADBOLT et al., 2012), permitindo que diversos setores possam compartilhar e interoperar dados. Dados Conectados podem ser vistos sob duas perspectivas: publicação e consumo. A publicação está sob a perspectiva do publicador, abordando conceitos (BERNERS-LEE, 2006; WOOD et al., 2014) e processos (BIZER et al., 2007; HYLAND; WOOD, 2011; VILLAZÓN-TERRAZAS et al., 2011; ÁVILA, 2015) necessários para publicar e manter os dados na Web de forma conectada. O consumo aborda o ponto de vista do consumidor, tratando a exploração de dados para tornar as aplicações mais ricas.

---

1 Para mais informações, acesse: <https://www.w3.org/standards/semanticweb/data>

2 A seriaização RDF é uma maneira de estruturar dados seguindo a estrutura de triplas. Algumas das serializações RDF são: RDF/XML, Turtle e N-triples. Mais informações no Capítulo de fundamentação.

3 RDFa é uma serialização RDF que utiliza atributos em documentos HTML. Para mais informações acesse: <https://rdfa.info>

4 json-ld é uma serialização RDF que utiliza o json como base, esta serialização é ideal para desenvolvedores, podendo ser utilizada em serviços Web e bancos de dados não relacionais (e.g. MongoDB). Para mais informações acesse: <http://json-ld.org>

Publicar ou manter dados conectados na Web vai além de disponibilizar conjuntos de dados através de serializações RDF. É necessário conectá-los a outros conjuntos de dados já existentes. Porém, criar *links* entre conjuntos de dados requer uma análise cuidadosa por parte do especialista, que apesar de ser uma abordagem eficaz, não é escalável, visto que a quantidade de dados publicados cresce constantemente. Consequentemente, inviabiliza o processo de publicação de forma manual. Logo, para que seja possível construir a Web de Dados de forma eficiente, é necessário que existam soluções capazes de conectar dados de forma automática ou semiautomática.

Problemática e justificativa

Conectar dados automaticamente é um problema reconhecido por diversas comunidades. Dentre essas comunidades, podemos citar a comunidade de Bancos de Dados e suas subáreas. Em Banco de Dados, esse problema é conhecido através do termo ***Record Linkage*** (GU et al., 2003), que tem como objetivo identificar e conectar recursos julgados representar a mesma entidade do mundo real. Além disso, é possível encontrar outras referências para esse problema, tais como: **Problema de Resolução de Entidades** (MENESTRINA; BENJELLOUN; GARCIA-MOLINA, 2005), **Deduplicação** (SARAWAGI; BHAMIDIPATY, 2002) e ***Instance Matching***.

Instance Matching se trata do termo que a comunidade de Dados Conectados utiliza para referenciar o problema. Nesta comunidade, o principal objetivo refere-se a encontrar instâncias correspondentes em *datasets* diferentes. Assim como em outras subáreas de Banco de Dados, existe uma interseção entre as técnicas e características. Porém, Castano et al. (2011) ressaltam que a correspondência de instâncias (*Instance Matching*) apresenta características adicionais tais como: (i) **heterogeneidade estrutural**, que se refere à variação na estrutura das instâncias; (ii) **conhecimento implícito**, que se refere às características e restrições apresentadas pelo domínio e (iii) **identificação orientada à URI**, refere-se ao reuso das URIs para identificar novas informações a respeito de instâncias já existentes. Logo, há a necessidade de soluções específicas para a execução correta do processo de correspondência de instâncias.

De acordo com (CASTANO et al., 2011), existem diversas técnicas que podem ser utilizadas no processo de correspondência de instâncias. Algumas dessas técnicas foram adaptadas de abordagens existentes na comunidade de Banco de Dados. Como pode ser visto na Figura 1, tais técnicas estão concentradas em duas categorias: a primeira se refere a abordagens orientadas a valor. Nesta categoria, assume-se que a similaridade entre recursos pode ser obtida através da correspondência entre os atributos desses recursos, vale a pena ressaltar que a maior parte das abordagens dessa categoria estão focadas na comparação entre textos (e.g. Dice, Levenshtein, Jaro etc.).

Na segunda categoria, estão concentradas as técnicas orientadas à instância.

Baseadas em aprendizagem: Utiliza grupos de treinamento bem como técnicas de

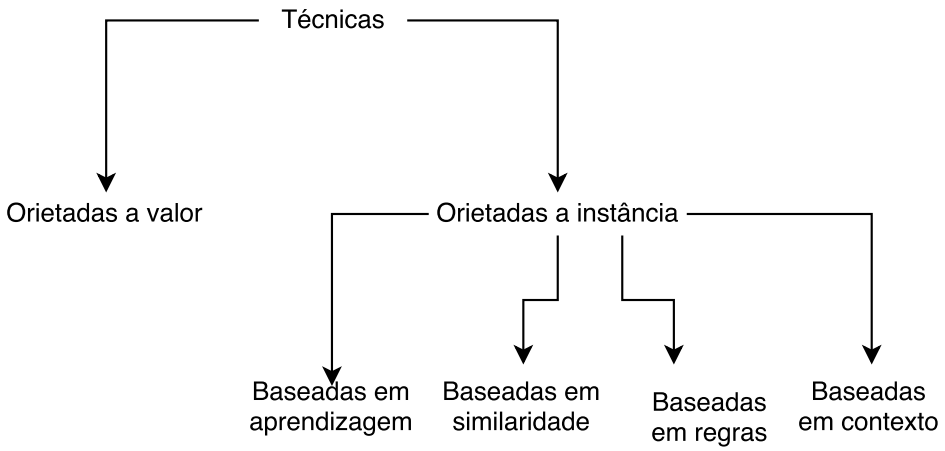


Figura 1 – Técnicas para correspondência de instâncias
Fonte: adaptado de (CASTANO et al., 2011).

aprendizagem de máquina, como Máquina de Vetor de Suporte (*support vector machine*), para definir se os recursos representam a mesma entidade do mundo real.

Baseadas em similaridade: Essa técnica enxerga os recursos como um conjunto de valores. Pode utilizar as mesmas funções para comparação entre textos, como a similaridade média para comparar dois recursos.

Baseadas em regras: Diferente das outras técnicas, esta usa valores *booleanos* no lugar de valores numéricos. Além disso, esta subcategoria apresenta bons resultados, embora seja dependente do domínio.

Baseadas em contexto: O contexto de um recurso pode ser entendido como os relacionamentos dele com outros recursos. Dessa forma, essa abordagem analisa as instâncias e suas relações.

Conforme as técnicas apresentadas, tem-se que o cálculo da similaridade entre recursos é elemento comum entre as aplicações de correspondência de instância.

Para identificar e conectar recursos na Web, a comunidade vem apresentando um número crescente de soluções (ver Figura 2). A *Ontology Alignment Evaluation Initiative* (OAEI) realiza uma avaliação anual, que consiste em alinhar dois conjuntos de dados pré-definidos e comparar o alinhamento gerado pela solução com o alinhamento de referência. A partir da comparação entre os dois conjuntos de alinhamento, as métricas de precisão, revocação e medida-f são geradas.

Porém, de acordo com Homoceanu, Kalo e Balke (2014), apesar dos bons resultados apresentados, as soluções não estão prontas para alinhar dados automaticamente de forma confiável. Para se chegar a esta conclusão, os autores realizaram um experimento equivalente ao executado pela OAEI, mas utilizando dados reais, que foram obtidos de 5 fontes diferentes

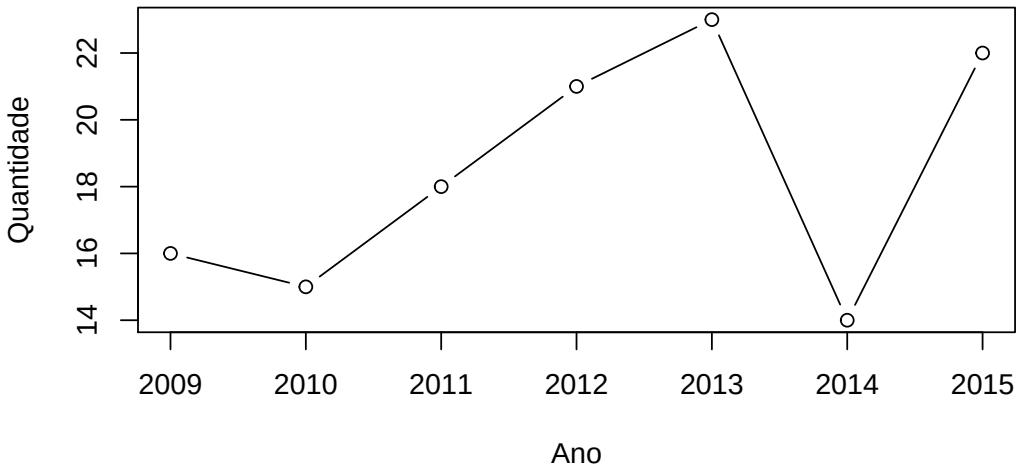


Figura 2 – Quantidade de soluções submetidas ao OAEI
Fonte: (CHEATHAM et al., 2015).

(Freebase[5], DBpedia[6], LinkedMDB[7], DrugBase[8] e NewYork Times[9]), que podem ser obtidos através do endereço <http://www.ifis.cs.tu-bs.de/node/2906>.

No experimento realizado por Homoceanu, Kalo e Balke (2014) foi utilizada uma abordagem de caixa preta, onde qualquer sistema independente de domínio pode ser utilizado na avaliação. Nesse contexto, o SLINT+ (NGUYEN; ICHISE; LE, 2012) foi utilizado por ser uma ferramenta independente de domínio e que não precisa de treinamento. A ferramenta foi avaliada em duas perspectivas. Na primeira perspectiva a transitividade da propriedade *owl:sameAs* foi desconsiderada. Diferentemente da segunda, que considerou a transitividade das correspondências criadas. A Tabela 1 apresenta os resultados obtidos no experimento (HOMOCEANU; KALO; BALKE, 2014).

Segundo (HOMOCEANU; KALO; BALKE, 2014) e (FERRARA et al., 2008), uma solução de correspondência de instâncias deveria explorar a modelagem ontológica, que utiliza vocabulários e ontologias para seu desenvolvimento. Essa recomendação se deve ao fato de RDF apenas prover estrutura aos dados, sendo responsabilidade das ontologias dar significado aos conceitos representados. Um modelo ontológico é composto por classes e propriedades, que são descritas através de uma linguagem de descrição de ontologias (e.g. OWL). As classes são utilizadas para representar os conceitos que pertencem ao domínio de interesse, já as propriedades são utilizadas para relacionar conceitos, sendo chamadas de propriedades de objeto

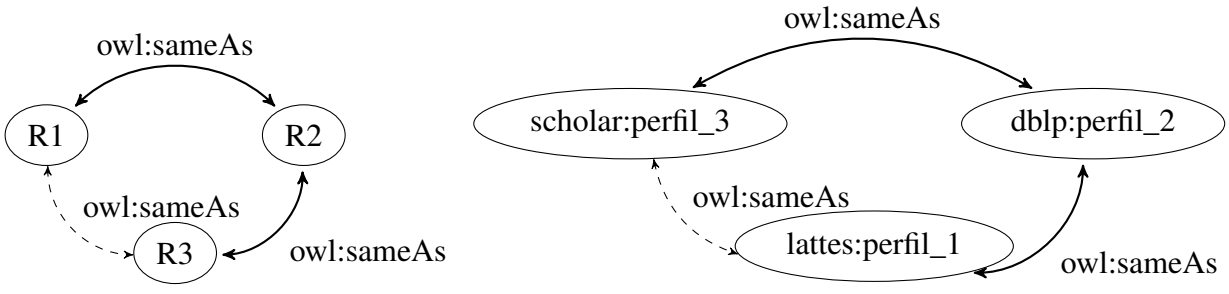
[5] <http://www.freebase.com/>
[6] <http://dbpedia.org>
[7] <http://www.linkedmdb.org>
[8] <http://www.drugbase.de/de/>
[9] <http://www.nytimes.com>

Tabela 1 – Quantidade de links *owl:sameAs*, quantidade de links *owl:sameAs* entre tipos diferentes, precisão. Nas perspectivas sem e com transividade para cada um dos limiares de aceitação *tetha*.

θ	SLINT+			cl_{TR}		
	#sameAs	Inter-domínio	Precisão	#sameAs	Inter-domínio	Precisão
0.95	8,020	33	0.91	2,055	89	0.20
0.75	16,739	119	0.71	5,498	216	0.15
0.50	17,436	230	0.76	7,038	396	0.09
0.25	25,113	1,734	0.67	14,879	2,408	0.02

(*object properties*) ou relacionar conceitos e dados, sendo chamadas de propriedade de dados (*data properties*). Essas propriedades possuem características como transitividade, simetria e reflexibilidade. Dessa forma, o baixo suporte às características das propriedades pode afetar diretamente na qualidade de soluções de correspondência de instâncias.

Dentre as propriedades com suporte inadequado, destaca-se a propriedade *owl:sameAs*, que é responsável por identificar recursos equivalentes. Além disso, essa se trata de uma propriedade transitiva, de forma que se existem dois recursos equivalentes R1 e R2 e existe um terceiro recurso R3 que é equivalente a R2, então R1 é equivalente a R3 como representado na Figura 3. Tais características fazem com que a propriedade *owl:sameAs* seja uma das mais utilizadas para alinhar dados na Web. Dessa forma, utilizar ferramentas que levem em consideração as características das propriedades é de grande importância para alinhar dados de forma confiável.



(a) owl:sameAs (b) Exemplo

Figura 3 – Transitividade da propriedade *owl:sameAs*

Neste contexto, este trabalho propõe uma abordagem independente de contexto para o alinhamento de dados conectados por meio de um processo de alinhamento que leva em consideração aspectos dos dados e características do modelo ontológico. Assim, os recursos/instâncias analisados, além de alinhados através das propriedades de dados, podem ser alinhados através de seus relacionamentos. Para isso, propõe-se uma abordagem de alinhamento em cascata. Ademais, a proposta trata o problema do alinhamento entre *datasets* reais, permitindo que seja possível alinhar *datasets* distribuídos na Web de forma confiável.

Em um cenário favorável, a tarefa de encontrar recursos correspondentes entre *datasets*

pode ser realizada com facilidade. Por outro lado, o mesmo não pode ser dito quando há milhares de instâncias, visto que os recursos necessários para manter uma análise manual tornam esta prática inviável. Além disso, devido ao esforço necessário juntamente com a quantidade de instâncias que devem ser comparadas, tem-se que a qualidade da análise pode ser comprometida.

Sabendo que aumentar a quantidade de recursos não é uma solução, nos deparamos com nosso problema geral:

(Problema Geral) Como simplificar a identificação de instâncias que dizem respeito à mesma entidade do mundo real?

Para manter a qualidade das análises, mesmo com o crescimento dos *datasets*, diversas comunidades utilizam a tecnologia para auxiliar na correspondência de recursos. Dentre essas comunidades, destaca-se a de Banco de Dados (BD), que estuda abordagens para identificar recursos entre bancos de dados diferentes. No entanto, Araujo et al. (2011) ressaltam que apesar das abordagens utilizadas pela comunidade de BD (*Record Linkage*) compartilhar características com as abordagens utilizadas pela comunidade de Dados Conectados (*Instance Matching*), ambas diferenciam-se em alguns aspectos, tais como: semântica dos relacionamentos e flexibilidade do modelo RDF.

Para suportar os aspectos que diferenciam *Record Linkage* de *Instance Matching*, a comunidade de Dados Conectados vêm desenvolvendo abordagens próprias. Essas abordagens vêm se mostrando promissoras diante das avaliações às quais são submetidas. Porém, apesar do desempenho apresentado, diversas soluções provêm suporte insuficiente a aspectos relacionados à semântica do modelo ontológico, comprometendo a qualidade das correspondências geradas.

O que leva ao nosso problema específico.

(Problema Específico) Como melhorar a eficácia das ferramentas para correspondência de instâncias?

Para tanto, é apresentado como questão de pesquisa (**Q1**) o seguinte questionamento:

Q1: Como se comportam as ferramentas para correspondência de instâncias com relação a eficácia?

Objetivo

Essa abordagem visa disponibilizar um mecanismo útil que permita alinhar recursos entre *datasets* diferentes. Além disso, a proposta também pretende facilitar a identificação e alinhamento de recursos dentro do mesmo *dataset*.

Apesar de ser um trabalho com enfoque em engenharia de software, as suas contribuições estão mais voltadas para a área de Dados Conectados. Segue algumas dessas contribuições:

- Construção de um processo para a correspondência de instâncias que seja independente de contexto;
- Implementação de uma abordagem em cascata para a correspondência de instância a partir de instâncias relacionadas;

Relevância e contribuições do trabalho

Anualmente a OAEI realiza a avaliação de ferramentas para a correspondência de instâncias. Essa avaliação utiliza *datasets* previamente disponibilizados juntamente com uma referência de correspondências entre as instâncias. Essa prática permite que o experimento seja reproduzido, possibilitando a validação dos resultados fornecidos pela OAEI. Por outro lado, permite também que desenvolvedores implementem algoritmos que utilizam computação específica para o *dataset*. Desta forma, é necessário o estudo de abordagens independentes de contexto que sejam capazes de alinhar dados de forma confiável. Além disso, as soluções fornecidas requerem que os *datasets* sejam fornecidos em formato de arquivo, sendo necessário gerar os arquivos dos *datasets* sempre que é necessário realizar a correspondência de instâncias.

Nesse contexto, a proposta apresenta as seguintes contribuições:

- Desenvolvimento de processo independente de contexto para o alinhamento de dados conectados;
- Viabilização da execução do alinhamento diretamente no armazenamento dos dados;
- Criação de experimento e estudo de caso para avaliar a eficácia das soluções de alinhamento no estado da arte.

Estrutura do trabalho

Esta dissertação está dividida em 6 capítulos. O Capítulo 2, são apresentados os conceitos e trabalhos relacionados ao ao trabalho proposto. Na conclusão do capítulo é apresentada uma tabela comparativa entre a abordagem proposta e os trabalhos relacionados apresentados.

O Capítulo 3 descreve em detalhes o processo e a arquitetura desenvolvida para suportá-lo. Estes, são apresentados através diagramas (atividades e componentes). Além disso, este capítulo conta também com uma descrição para cada etapa do processo.

No Capítulo 4, um experimento foi projetado para avaliar, em termos de eficácia através das métricas de precisão, revocação e medida-f, a abordagem proposta, em comparação com AgreementMakerLite (AML) (FARIA et al., 2016) e RiMOM-2016 (ZHANG et al., 2016). Cada conjunto de alinhamento gerado é avaliado e uma discussão geral é apresentada ao final do capítulo.

No Capítulo 5 é apresentado um estudo de caso. Nele é descrito o QIE, um sistema que apresenta o cruzamento dados da Revista Brasileira de Informática na Educação (RBIE), Simpósio Brasileiro de Informática na Educação (SBIE) e Workshop de informática na Escola (WIE) com dados extraídos da plataforma LATTES. Além disso, é descrito como o processo foi utilizado para alinhar os dados dos pesquisadores e de suas produções científicas entre essas bases.

Por fim, no Capítulo 6, são apresentadas as considerações finais deste trabalho.

2 FUNDAMENTAÇÃO E TRABALHOS RELACIONADOS

Este capítulo está estruturado em duas seções, a primeira apresenta os conceitos necessários para o entendimento desse trabalho. A segunda apresenta os trabalhos que estão relacionados à correspondência de instância.

Fundamentação Teórica

O objetivo desta seção é apresentar a fundamentação teórica referente ao foco deste trabalho, fazendo uma pequena introdução sobre conceitos da área que auxiliam na compreensão da pesquisa desenvolvida.

RDF (Resource Description Framework)

RDF trata-se de um *framework* de descrição de recursos, trabalhando como um alicerce para a construção de Dados Conectados. Em concordância com tecnologias Web, bem como Dados Conectados, o modelo RDF utiliza URIs para identificação de recursos, permitindo que os recursos sejam descritos uniformemente na Web. Basicamente, o RDF é estruturado em triplas do tipo sujeito, predicado, objeto (ver Figura 4).

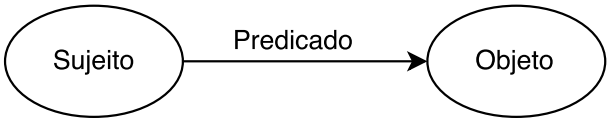


Figura 4 – Estrutura da tripla RDF

A estrutura do RDF pode se materializar de várias formas. Cada uma dessas formas recebe o nome de serialização RDF. Atualmente existe um conjunto considerável de serializações tais como: RDF/XML, Turtle, N-Triples e outros. Onde cada serialização tem um uso em potencial. Por exemplo, Turtle é usado para ser lido por humanos, pois ele é melhor estruturado para isso.

Vale a pena salientar que RDF é um framework de descrição, não sendo responsável por atribuir semântica aos recursos descritos, sendo as ontologias responsáveis por esta função. Por esta razão, um número considerável de processos de publicação de Dados Conectados (BIZER et al., 2007; HYLAND; WOOD, 2011; VILLAZÓN-TERRAZAS et al., 2011; ÁVILA, 2015) recomenda o reuso de ontologias.

Ontologias

Os Dados Conectados fazem uso de ontologias como suporte formal para representação de conhecimento, pois só RDF não é o bastante para que máquinas consigam entender a relação entre os dados. Desta forma, as ontologias têm um papel fundamental na modelagem e descrição de dados. A palavra ontologia vem do Grego *ontos* e *logos*, significando conhecimento do ser. Em filosofia, ontologia refere-se ao estudo do ser. Em Computação, de maneira informal, uma ontologia define um conjunto de conceitos e suas relações, tais como a terminologia (vocabulário do domínio), definição explícita dos conceitos essenciais, suas classificações, taxonomias, relações e axiomas do domínio, incluindo hierarquias e restrições (DEVED et al., 2006). Em 1993, GRUBER et al. definiu ontologia como uma especificação explícita de uma conceitualização. Em 1997, BORST definiu ontologia como uma especificação formal de uma conceitualização compartilhada. Em 1998, STUDER et al. unificaram as duas definições, desta forma ontologia pode ser definida como uma especificação formal e explícita de uma conceitualização compartilhada. Para melhor entendimento, abaixo seguem detalhes sobre os termos citados nessa definição:

- **Explícita:** definições de conceitos, relações, restrições e axiomas;
- **Formal:** compreensível para agentes e sistemas;
- **Conceitualização:** modelo abstrato de uma área de conhecimento;
- **Compartilhada:** conhecimento consensual.

Pode-se concluir, a partir das características supracitadas sobre ontologias, que o conhecimento é explícito. Esse conhecimento equivale à descrição de determinada área do conhecimento, garantindo um conhecimento "consensual" sobre tal área. A partir do momento em que há um consenso, há a possibilidade e a viabilidade de compartilhar tais ontologias e integrá-las a outras áreas de conhecimento (através de outras ontologias).

Dados Conectados

O termo Dados Conectados (do inglês Linked Data) é a tradução oficial para o conceito na língua portuguesa (ISOTANI; BITTENCOURT, 2015). Segundo o W3C, Dados Conectados pode ser entendido como o núcleo da Web Semântica, tendo como objetivo prover a integração e raciocínio de dados disponíveis na Web. Além disso, Berners-Lee (2006) ressalta que Dados Conectados não se trata apenas de pôr dados na internet, mas fazer conexões entre eles, permitindo que pessoas ou máquinas possam explorar a Web dos dados.

Para conectar os dados disponíveis na Web com qualidade, foram desenvolvidas as boas práticas. Essas boas práticas são fundamentadas em tecnologias Web como ressaltam Isotani e

Bittencourt (2015). Além dessas tecnologias, vale a pena destacar os quatro princípios básicos de Dados Conectados (BERNERS-LEE, 2006):

1. Usar URI para a identificação de recursos
2. Usar HTTP URIs para que seja possível buscar pelos recursos
3. Prover informação útil para as URIs consultadas através de padrões (RDF e SPARQL)
4. Incluir links para outras URIs. Possibilitando a descoberta de novos recursos

É possível dividir os princípios em duas categorias. A primeira categoria é composta pelos dois primeiros princípios, estando relacionada a identificação e resolução desses recursos através de URI e HTTP URIs. A segunda categoria está relacionada de forma prática à conexão dos dados, utilizando RDF para especificar como os recursos são descritos e URIs que apontam para outros recursos, conectando de fato os dados.

Algoritmos de similaridade

Algoritmos de similaridade podem ser entendidos como funções utilizadas para medir a semelhança entre objetos. Essas funções podem ser utilizadas em diferentes contextos, partindo de correções ortográficas até tarefas de processamento de linguagem natural (PLN). A escolha de uma abordagem pode variar de acordo com o contexto de aplicação. Além disso, existe mais de uma abordagem dentro do mesmo contexto.

A Figura 5 apresenta uma visão geral sobre as abordagens de acordo com seu contexto de aplicação (tracejado). Esses contextos estão descritos a seguir:

- **Baseadas em edição:**
  As técnicas baseadas em edição calculam a similaridade entre objetos através do número de inserções e remoções necessárias para que um objeto se torne igual ao outro. Dentre as abordagens existentes, tem-se que esta é uma das mais conhecidas. Alguns fatores que corroboram para a sua popularidade são a simplicidade e que um dos algoritmos de similaridade mais conhecidos pela comunidade, algoritmo de Levenshtein (LEVENSHTEIN, 1966). Esse algoritmo foi proposto em 1965 e se baseia na quantidade mínima de edições (remoções e adições) necessárias para que uma palavra seja igual a outra.

- **Baseadas em token:**
  As técnicas baseadas em token consideram que cada palavra é um elemento de conjunto, que pode ser ordenado (vetor) ou não ordenado. Neste contexto, tem-se que o objetivo das técnicas é analisar quão similar são estes conjuntos. Na perspectiva de conjunto como vetor, tem-se que quanto menor é o ângulo existente entre os vetores mais similares eles

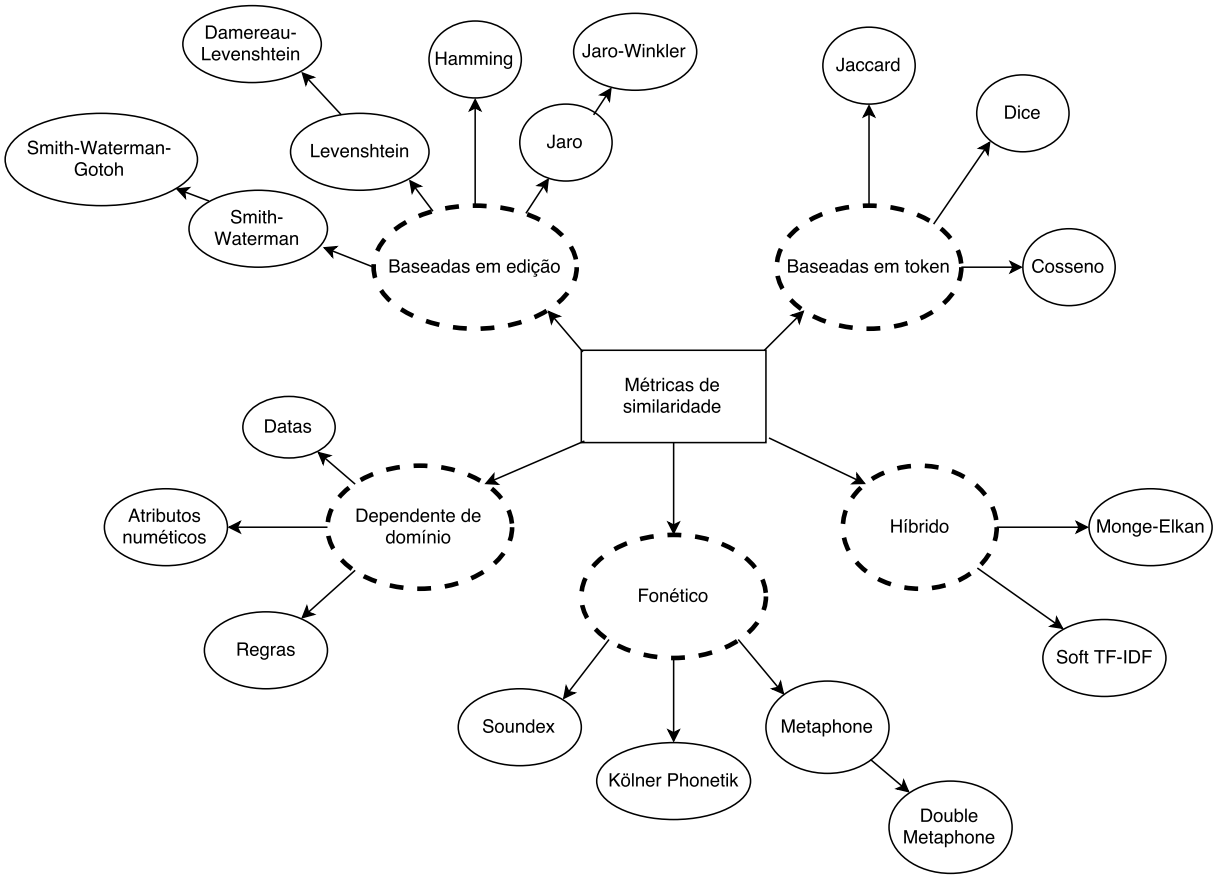


Figura 5 – Visão geral das técnicas de similaridade

são. Na perspectiva de conjunto, a similaridade entre os conjuntos é dada pela quantidade de elementos comuns aos dois conjuntos.

- **Dependentes de domínio:**
  As técnicas dependentes de domínio exploram informações relacionadas aos dados. Essas informações são de conhecimento prévio do especialista. Dados como: datas, valores e regras de associação, são utilizados para analisar a similaridade entre objetos. Vale ressaltar que as técnicas baseadas em contexto possuem baixo índice de reuso.

- **Fonéticas:**
  As técnicas baseadas em similaridade fonética exploram a semelhança existente entre os sons. Isso quer dizer que palavras homófonas apresentam o mesmo valor. Por exemplo, o valor gerado pela função *soundex* para as entradas **xeque** e **cheque** são equivalentes (("22).

- **Híbridas:**
  As técnicas baseadas em abordagem híbrida, por sua vez, permitem que outras técnicas de similaridades sejam compostas entre si. Por exemplo, a técnica denominada Monge-Elkan (MONGE et al., 1996), que permite a utilização de outras técnicas internamente.

Esta técnica executa em ordem quadrática, visto que a similaridade escolhida por cada componente é dada pela maior similaridade encontrada durante as comparações.

Nesta proposta, foi utilizada uma generalização do algoritmo de Monge-Elkan (MONGE et al., 1996), que tem o objetivo de dar um peso maior a *tokens* mais semelhantes. Inicialmente, Monge-Elkan foi desenvolvido com o objetivo de calcular a semelhança entre textos que possuem vários *tokens*, onde a similaridade de cada *token* é calculada através da média das similaridades internas, que é provida por uma técnica baseada em caractere (e.g. Cosseno, Jaro, Dice). Essa abordagem se destaca em cenários de desordem ou ausência de *tokens*. Além disso, esse algoritmo explora os benefícios providos pelas abordagens baseadas caractere (e.g. erros de digitação, erros de OCR e erros ortográficos) (JIMENEZ et al., 2009). Por fim, a definição formal é descrita a seguir:

$$Sim_{MongeElkan}(x, y) \quad = \quad \frac{1}{|x|} \sum_{i=1}^{|x|} \underset{j=1,|y|}{max} sim'(x\,[i]\,, y\,[j]) \tag{2.1}$$

Onde:

- $|x|$ é q quantidade de tokens em x
- sim' é uma função de similaridade interna (e.g., Jaro-Winkler)

Como se sabe, o algoritmo de Monge-Elkan utiliza permite que outras funções de similaridade sejam utilizadas internamento. Desta forma utilizaremos o algoritmo Jaro-Winkler pra exemplificar sua utilização. Inicialmente, imaginemos as seguintes entradas "Armando Barbosa"e "Barbosa A". Cada entrada é considerada um conjunto, dessa forma cada palavra é considerada separadamente. A seguir, as palavras de um conjunto são comparadas com as do outro buscando os pares com maior pontualção. Essa pontuação é calculada através da função de similaride interna, neste caso, o algoritmo de levenshtein, obtendo os seguintes resultados.

jaro_winkler("Armando","Barbosa") = 0.50
jaro_winkler("Armando","A") = 0.74
jaro_winkler("Barbosa","Barbosa") = 1
jaro_winkler("Barbosa","A") = 0

Dessa forma, temos:

monge_elkan = $\frac{(0.74+1)}{2} = 0.87$

Alinhamento de Dados Conectados

Alinhar Dados Conectados trata-se de um processo que tem como objetivo identificar e mesclar recursos que representam a mesma entidade do mundo real. Segundo Homoceanu, Kalo e Balke (2014), no contexto de dados conectados temos a seguinte definição para o problema:

$$f\left(URI_i, URI_j\right) := \begin{cases} \text{verdadeiro, se } sim\left(URI_i, URI_j\right) > \theta \\ \text{falso, caso contrário} \end{cases} \quad \text{com } URI_i \in D_i \text{ e } URI_j \in D_j \tag{2.2}$$

onde $1 \leq$ i,j $\leq$ n, e $sim()$ é uma função que é capaz de calcular a similaridade entre os recursos. Esses recursos pertencem a *Datasets* distintos(D_i e D_j). Por fim, tem-se que θ é o parâmetro que regula o nível de qualidade para o alinhamento, que também pode ser conhecido como limiar ou *threshold*. Além disso, é possível ressaltar outros benefícios que surgem a partir da conexão entre dados, sendo elas:

- **Integração semântica de dados:** Refere-se ao melhoramento das técnicas existentes para descobertas de mapeamento (semi) automático entre ontologias heterogêneas e distribuídas;
- **Reconhecimento de identidade:** Refere-se à capacidade de identificar se descritores de recursos distintos estão relacionados à mesma entidade do mundo real;
- **População de ontologias:** Refere-se à descoberta de relacionamentos entre novas instâncias e as instâncias já existentes na base de conhecimento.

Segundo Ferrara et al. (2008), para que uma abordagem seja capaz de identificar recursos que identifiquem a mesma entidade do mundo real, essa deve satisfazer diferentes requisitos, que estão dispostos em três categorias (ver Figura 6).

Valores diferentes: Um algoritmo de correspondência de instância deve reconhecer valores correspondentes, sempre que possível, mesmo quando esses valores possuem erros. Para mitigar esse problema, a comunidade utiliza abordagens como algoritmos de similaridade e transformação de valores.

Heterogeneidade Estrutural: Instâncias que pertencem a ontologias diferentes diferem não somente entre propriedades e valores, mas também na sua estrutura. Desta forma, algoritmos de alinhamento de dados devem identificar propriedades semelhantes em ambos os recursos.

Heterogeneidade Lógica: A heterogeneidade lógica trata-se de um problema de alinhamento de ontologias, o qual não é levado em consideração no processo de alinhamento de dados, no entanto, faz-se necessário em tarefas de inferência. Esse problema diz respeito à semântica

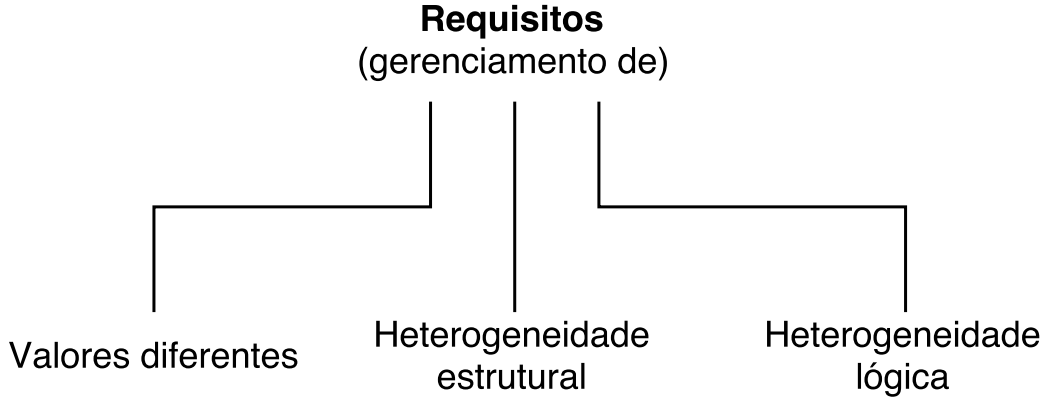


Figura 6 – Requisitos para soluções de alinhamento de dados
Fonte: baseado em (FERRARA et al., 2008)

atribuída aos termos, havendo situações em que o mesmo termo pode ter significados diferentes atribuídos a ele (*e.g.* manga - fruta e manga - vestimenta).

Diante do exposto, a comunidade vem desenvolvendo alternativas para a identificação de correspondência de instâncias. Diante disso, a seção a seguir apresenta os trabalhos relacionados a esta proposta.

Trabalhos Relacionados

Neste capítulo serão apresentadas duas ferramentas para o alinhamento de dados conectados. As ferramentas apresentadas a seguir foram selecionadas devido ao seu destaque na edição de 2016 do relatório publicado pela Ontology Alignment Evaluation Initiative (OAEI), mais especificamente na trilha referente à correspondência de instâncias (*Instance Matching*). Inicialmente, a OAEI avaliava apenas ferramentas de alinhamento de ontologias, dando início à avaliação de soluções para alinhar dados em 2009. Desde então, um número crescente aplicações vêm sendo submetidas (CHEATHAM et al., 2015).

AgreementMakerLight (AML)

Desenvolvido em parceria entre o Instituto Gulbenkian de Ciência, a Universidade de Lisboa, e a Universidade de Illinois, o AgreementMakerLight é uma ferramenta de alinhamento de ontologias. De acordo com (FARIA et al., 2016), o AML se baseia inicialmente em técnicas e similaridade léxica, tendo como ênfase o uso de fontes externas como *background*.

O AML conta com três algoritmos para alinhamento voltados a correspondência de instâncias, sendo eles o *HybridStringMatcher*, o *ValueStringMatcher* e o *Value2LexiconMatcher*. O primeiro utiliza diversas abordagens para gerar a similaridade, sendo elas a comparação entre frases, palavras. Além disso, essa abordagem híbrida também explora a *WordNet*. O

segundo utiliza o mapeamento de valor para calcular a similaridade, penalizando pares nos quais anotações ou propriedades de dados não são os mesmos. Por fim, o terceiro une as duas abordagens anteriores.

Apesar do AML possuir diferentes algoritmos de alinhamento na ferramenta, todos eles trabalham apenas no nível dos dados. Consequentemente, as características das propriedades são desconsideradas ao longo do processo de correspondência.

RiMOM-2016

Baseando-se no RiMOM (LI et al., 2009), Zhang et al. (2016) desenvolveram o RiMOM-2016, que é uma ferramenta para alinhar dados conectados. Ela implementa um número considerável de abordagens para alinhar, cuja escolha é realizada através dos metadados extraídos da ontologia. Além disso, o RiMOM-2016 utiliza um índice invertido para indexar os objetos e consequentemente gerar pares candidatos para um possível alinhamento. A geração dos pares é realizada quando dois recursos compartilham pelo menos um predicado e objeto.

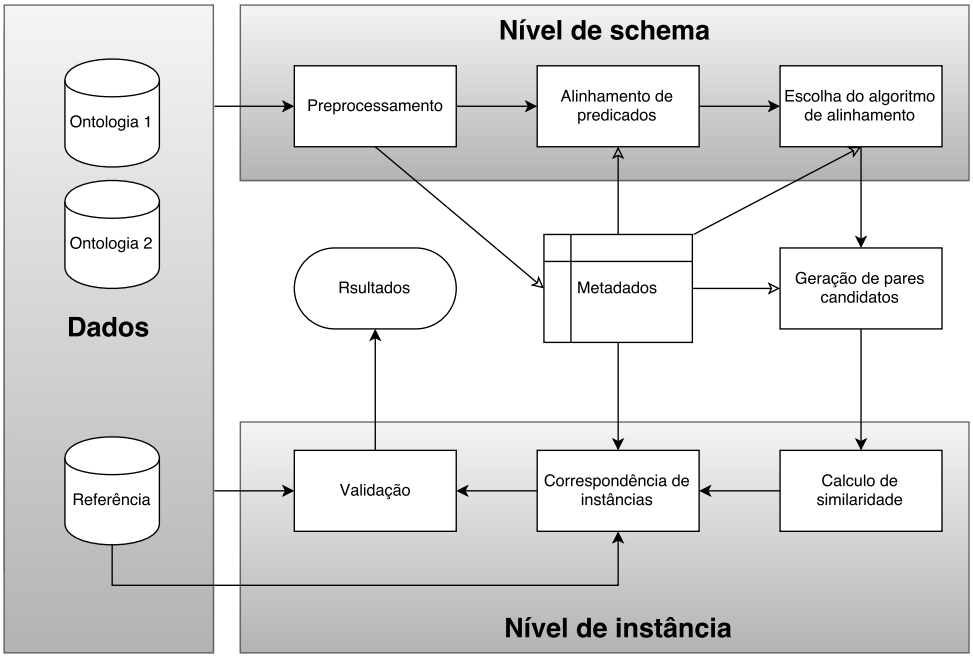


Figura 7 – Arquitetura do RiMOM-2016
Fonte: adaptado de (ZHANG et al., 2016)

Por um lado, o índice invertido permite que um número menor de comparações seja realizado. Por outro lado, a etapa de construção desse índice não considera que os objetos indexados podem conter qualquer tipo de erro. Além disso, como pode ser visto na Figura 7, o RiMOM-2016 utiliza as ontologias apenas para alinhar as propriedades e como entrada para a geração de metadados.

Comparação com a proposta

Neste capítulo, algumas das principais ferramentas existentes foram apresentadas. Essas ferramentas tem o objetivo de alinhar dados conectados através de diversas abordagens. Dentre os sistemas apresentados, nenhum deles contempla o alinhamento de dados como foco principal, sendo variações de ferramentas existentes para o alinhamento de ontologias.

Esta seção tem o objetivo de apresentar um comparativo entre os trabalhos relacionados e o presente estudo. Para isso, consideramos os seguintes critérios:

1. **Quantidade de *datasets* suportados:** Este critério refere-se à quantidade de de *datasets* que podem ser utilizados para no processo de correspondência de instâncias. Normalmente, apenas dois *datasets* são suportados simultaneamente. Com isso, a troca de um dos *datasets* é necessária.

2. **Tipo de suporte à ontologia:** Este critério refere-se à maneira como as ontologias são exploradas pela ferramenta de correspondência de instâncias, podendo ser classificada como dirigida ou baseada em ontologia. Uma ferramenta pode ser classificada como dirigida por ontologia quando explora a modelagem conceitual com o objetivo de escolher o conjunto de atividades que são utilizados no processo. Por outro lado, uma ferramenta é dita baseada em ontologias quando a modelagem conceitual é a base para o processo de correspondência.

3. **Utilização de computação específica:** Este critério refere-se à utilização de algoritmos específicos para os *datasets* que serão utilizados no processo de correspondência de instâncias com o objetivo de melhorar a precisão das correspondências estabelecidas.

4. **Exploração dos conceitos presentes na ontologia:** Este critério refere-se a exploração dos conceitos presentes na modelagem conceitual. Além disso, as características dos relacionamentos são consideradas (*e.g.* transividade).

Tabela 2 – Comparação entre os trabalhos

Critérios de comparação	**RiMOM**	**AML**	**Proposta**
(1) Quantidade de *datasets* suportados	2	2	2+
(2) Tipo de suporte à ontologia	Dirigido	Dirigido	Baseado
(3) Utilização de computação específica	Sim	Sim	Não
(4) Exploração dos conceitos presentes na ontologia	Não	Não	Sim

Contudo, apesar das ferramentas se mostrarem capazes de encontrar correspondência entre instâncias, ainda deixam a desejar em alguns critérios, tais como a utilização de computação específica para os *datasets*, além da pouca exploração das ontologia, que são utilizadas apenas para a geração de metadados, com o objetivo de escolher entre as abordagens de correspondência

disponíveis. Diferentemente das ferramentas citadas, a proposta utiliza ontologias para guiar o processo de correspondência de instância. Além disso, essa abordagem permite que o usuário defina como o alinhamento deve ser realizado.

Outro diferencial da proposta em relação aos trabalhos relacionados está no que chamamos de alinhamento em cascata, que consiste na utilização de instâncias que pertencem a conceitos relacionados ao conceito cujas instâncias serão alinhadas. O alinhamento em cascata vai além das instâncias, ele explora os relacionamentos existentes. A partir disso, é possível encontrar novas correspondências. Ademais, a proposta permite que as correspondências entre as instâncias sejam armazenadas diretamente na base de triplas (*triple store*) em que estão armazenadas.

3 PROPOSTA

Processo de alinhamento

Neste capítulo, será apresentado o processo proposto para alinhar dados. Como exibido na Figura 8, o processo é composto por 4 etapas principais, sendo elas: selecionar *datasets*, identificar conceitos, listar recursos e alinhar dados. Cada etapa do processo será descrita nas subseções a seguir.

I - Selecionar Conjunto de Dados

A etapa de selecionar *datasets* visa determinar quais conjuntos de dados serão alinhados. A seleção de um *dataset* está sujeita a alguns critérios, tais como: estar estruturado em triplas e utilizar conceitos modelados em ontologias/vocabulários.

Tais critérios foram definidos de acordo com o escopo do processo, ou seja, dados conectados. Nesta área, há ferramentas e processos disponíveis para a publicação de dados conectados na Web, que apesar de estar relacionado ao trabalho proposto não pertence ao escopo do mesmo. Além disso, é importante destacar que ao modelar os dados em qualquer processo de publicação de dados conectados são utilizadas ontologias/vocabulários que podem servir como base. Complementarmente, além de se adequar ao escopo, os critérios cumprem os requisitos mínimos para a execução do processo.

II - Identificar Conceitos

Para auxiliar na escolha do conceito principal bem como os conceitos que estão relacionados, foram desenvolvidas duas consultas SPARQL. A primeira consulta explora a ontologia, principalmente as relações *rdfs:domain* e *rdfs:range* das propriedades de objeto (ver código 3.1). A segunda explora os dados e as relações estabelecidas pelas instâncias. Na consulta (código 3.1), a linha 4 tem o papel de recuperar todos os conceitos pertencentes à ontologia ou vocabulário. Na linha 5 é aplicada uma restrição, em que os conceitos devem ser domínio ou *range* de uma relação. Consequentemente, uma instância desse conceito será sujeito ou objeto de uma tripla (ver Figura 9).

A consulta apresentada no Código 3.2 é composta de duas partes, visto que o conceito pode modelar instâncias que são sujeito ou objeto de uma relação. Na primeira parte, o conceito selecionado representa o sujeito da tripla. Através das relações das instâncias é possível recuperar os conceitos que modelam as instâncias relacionadas (objetos). Na segunda parte ocorre o inverso, o conceito representa o objeto da tripla e os conceitos que representam os sujeitos são recuperados.

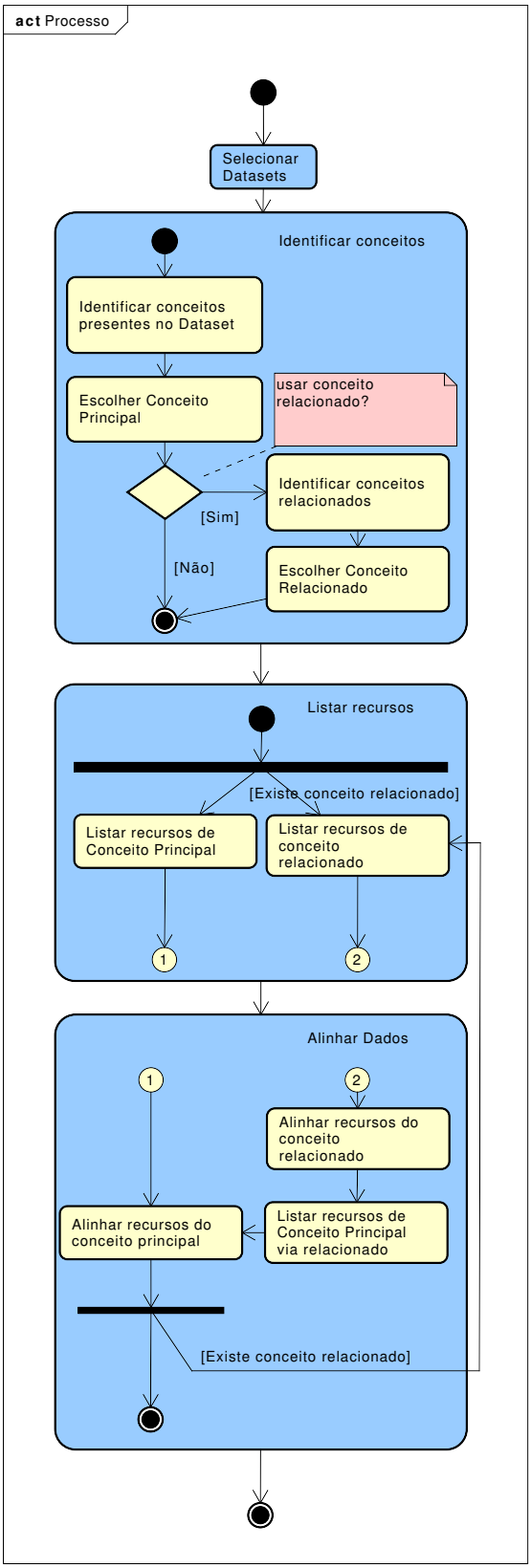


Figura 8 – Processo de alinhamento de dados conectados

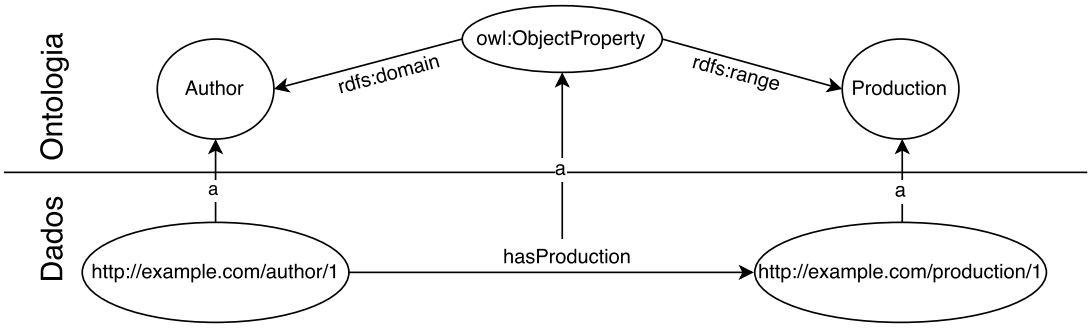


Figura 9 – Relação entre ontologia e dados

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
select distinct ?Concept count(*) as ?count where {
[]      a ?Concept.
?Concept        (^rdfs:domain|^rdfs:range) ?o.
}
group   by ?Concept
order   by desc(?count)
```

Código 3.1 – Consulta SPARQL para identificação de conceito

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>


select distinct ?type where {
        values ?Concept{<URI do conceito escolhido>}
        {
                ?instance rdf:type ?Concept; ?p ?o.
                ?p rdf:type owl:ObjectProperty.
                ?o rdf:type ?type.
        }
        union
        {
                ?s ?p ?o.
                ?p rdf:type owl:ObjectProperty.
                ?o rdf:type ?Concept.
                ?s rdf:type ?type.
        }
}
```

Código 3.2 – Query SPARQL para recuperação de conceitos relacionados

Como resultado do Código 3.2 é provida uma lista contendo os conceitos relacionados ao conceito (principal) escolhido. Neste momento, o usuário deve escolher, quais conceitos relacionados ele deseja utilizar para melhorar o alinhamento do conceito escolhido. A escolha dos conceitos, assim como a quantidade de conceitos relacionados pode ser realizada de forma arbitrária. Essa decisão influenciará tanto no tempo que o processo levará para concluir, quanto na quantidade de recursos alinhados ao final do processo, pois para cada conceito relacionado haverá uma nova execução das etapas (iii) e (iv). Esse *loop* é necessário, pois alguns alinhamentos serão possíveis apenas através da relação entre esses conceitos.

III - Listar Recursos

A etapa de listar recursos pode ser entendida como a recuperação dos recursos que pertencem aos conceitos. É importante destacar que a listagem/recuperação de recursos da base de conhecimento pode ser executada mais de uma vez durante o processo, gerando um conjunto de recursos para cada conceito escolhido. Além disso, essa etapa é responsável pela geração de pares candidatos, onde os recursos do *Dataset* D_1 são comparados com os recursos do *Dataset* D_2.

IV - Alinhar Dados

A etapa de alinhamento de dados está dividida em duas atividades sendo elas: (i) alinhamento simples e (ii) alinhamento em cascata, que serão detalhados nas subseções a seguir.

Alinhamento simples

Para alinhar os recursos é necessário executar alguns procedimentos, sendo eles o tratamento dos dados, comparação entre recursos e análise da similaridade. O primeiro procedimento, tratamento, se refere a transformações nas propriedades dos recursos. Essas transformações são necessárias para auxiliar os algoritmos de similaridade a analisarem melhor a semelhança entre os recursos. No procedimento de comparação, cada uma da propriedades é analisada. Caso uma propriedade não pertença a um dos recursos, ela é dispensada da comparação. A Figura 10 apresenta a comparação entre as propriedades de cada recurso.

Para definir a similaridade entre as instâncias foram usadas duas equações. A Equação 3.1 define o conjunto de propriedades que será considerado durante a comparação entre os recursos, que será obtido a partir da diferença entre o maior conjunto de propriedades e o conjunto de propriedades que deve ser desconsiderado. Logo:

$$P_f = Max(P_{r1}, P_{r2}) - P_d \tag{3.1}$$

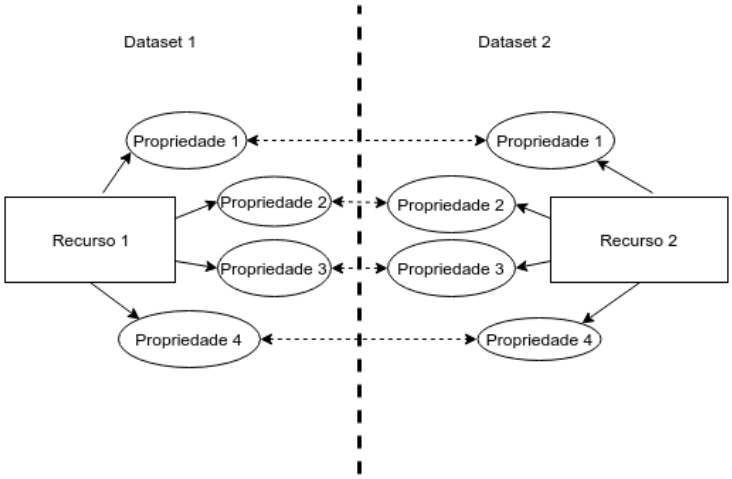


Figura 10 – Comparação entre recursos

Onde:

- P_{r1} – Conjunto de propriedades do recurso 1;
- P_{r2} – Conjunto de propriedades do recurso 2;
- P_d – Conjunto de propriedades que devem ser desconsideradas;
- $Max(P_{r1}, P_{r2})$ – Retorna o conjunto com maior número de propriedades;

A Equação 3.2 trata da função de similaridade entre recursos, essa equação pode ser entendida como a média das similaridades entre dois recursos. Tal abordagem foi escolhida com o objetivo de não privilegiar nenhuma das similaridades parciais. Contudo, é possível que existam outras funções mais adequadas para calcular a similaridade entre recursos.

$$SR = \frac{1}{|P_f|} \sum_{i=1}^{P_f} S(V(R_1, P_f[i]); V(R_2, P_f[i])) \tag{3.2}$$

Onde:

- S – Função de similaridade;
- V(R,P) – Valor da propriedade P em um recurso R;
- R_1 – Recurso 1;
- R_2 – Recurso 2;

O valor gerado pelo componente de similaridade é enviado para o componente de alinhamento.

Alinhamento em cascata

O alinhamento em cascata recebe esse nome devido ao encadeamento entre as atividades necessárias: (i) Recuperar instâncias que pertencem ao conceito relacionado; (ii) Alinhar instâncias que pertencem ao conceito relacionado; (iii) Recuperar instâncias que pertencem ao conceito principal; e (iv) Alinhar instâncias que pertencem ao conceito principal. O nome também foi utilizado com a intenção de fazer referência ao modelo de desenvolvimento em cascata (ROYCE, 1970), que é o primeiro modelo de desenvolvimento de software. Segundo Sommerville (2011), trata-se de um modelo dirigido por planos, cujas etapas são planejadas antes da execução.

O modelo de desenvolvimento em cascata e a abordagem de alinhamento em cascata compartilham algumas semelhanças. Dentre as semelhanças compartilhadas, temos o modo como as atividades são executadas, que é sequencial. Além disso, cada atividade só pode ser iniciada quando a atividade anterior for concluída. Outra característica compartilhada entre eles é o fato de todo o projeto ser planejado antes da execução.

Diferentemente de um projeto que utiliza o modelo em cascata, onde todo o projeto deve estar concluído na etapa final, tem-se que o processo de correspondência entre instâncias só será considerado concluído quando todos os conceitos relacionados forem utilizados no processo de correspondência. Vale ressaltar que uma "cascata" é gerada para cada conceito relacionado selecionado.

A Figura 11 apresenta o relacionamento entre os conceitos e a cascata.

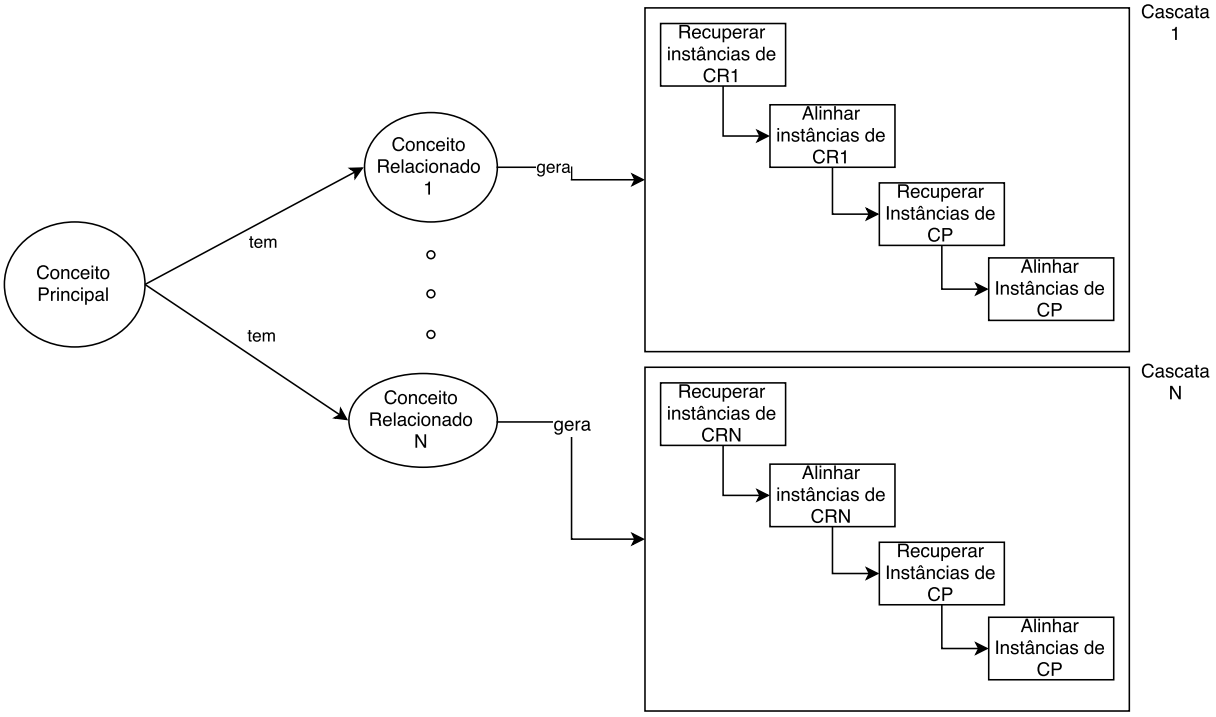


Figura 11 – Relacionamento entre conceito relacionado e cascata

Imaginemos que algum órgão deseja encontrar quais autores estão presentes em dois

datasets simultaneamente. Para isso, desejam utilizar os artigos presentes em ambas as bases. Dessa forma, temos que o conceito principal e conceito relacionado são ***Author*** e ***Publication***, respectivamente.

Recuperar Instâncias do Conceito Relacionado

Nesta atividade recuperam-se as instâncias que pertencem a um conceito relacionado. Para isso, consultas são realizadas nos *datasets* selecionados na primeira etapa do processo. Os resultados das buscas são agrupados em listas, uma para cada *dataset*. Cada elemento da lista é um grafo que é composto pela instância e seus relacionamentos. Por fim, as listas são enviadas para a próxima atividade da cascata, como pode ser visto na Figura 11.

Tomando como referência o exemplo citado, temos que as listas serão compostas por instâncias de ***Publication*** e seus relacionamentos.

Alinhar Instâncias do Conceito Relacionado

As instâncias recuperadas na atividade 3.1.4.2.1 são submetidas para o alinhamento. Através do cálculo de similaridade realizado entre essas instâncias. O resultado desse cálculo é comparado com um limiar de aceitação, havendo duas possibilidades. Em caso positivo, ou seja, a similaridade está no limiar de aceitação. Consequentemente, as instâncias são entendidas como correspondentes e o alinhamento é persistido. Caso contrário, o par candidato é descartado.

Recuperar Recursos do conceito principal

Similarmente ao que ocorre na subseção 3.1.4.2.1, a recuperação das instâncias que pertencem ao conceito principal (***Author***) também é realizada através de consultas nos *datasets* selecionados. Porém, neste caso, as instâncias recuperadas devem estar relacionadas às instâncias da correspondência persistida anteriormente. Dessa forma, os autores que estão relacionados à publicação são recuperados e agrupados de acordo com o *dataset*. A Figura 12 apresenta a relação entre as instâncias que pertencem ao conceito principal e as correspondências.

Alinhar Recursos Recuperados

Através da restrição aplicada na atividade 3.1.4.2.3, tem-se que as instâncias recuperadas estão relacionadas à mesma entidade do mundo real. Consequentemente, existe ao menos uma correspondência entre as instâncias recuperadas. Assim, a correspondência entre os autores é persistida.

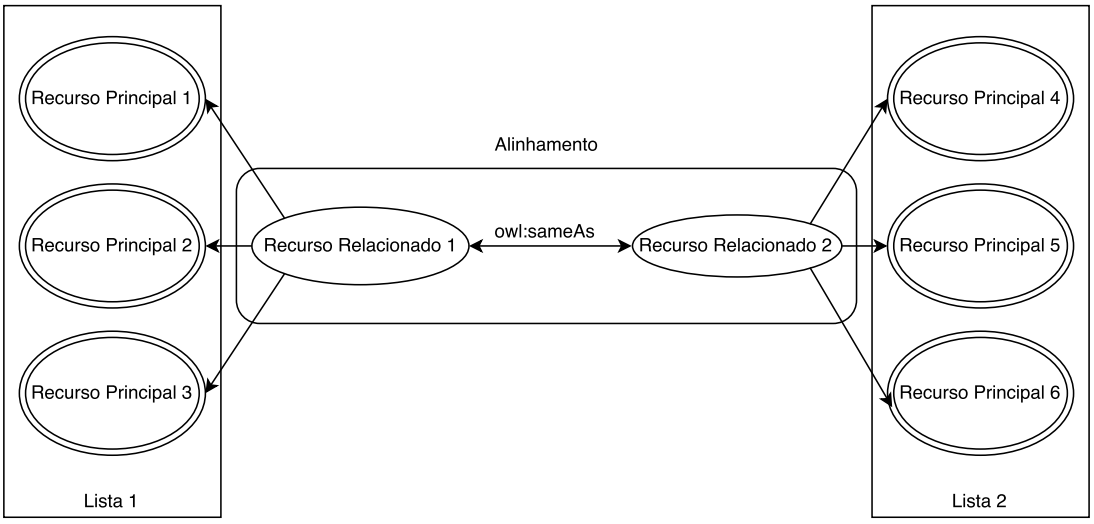


Figura 12 – Relacionamento entre instâncias e correspondência

Considerações Finais

Este capítulo apresentou um processo para o alinhamento de dados. Este processo foi projetado para permitir que soluções sejam capazes de estabelecer a correspondência entre instâncias dispensando a utilização de computação específica, em outras palavras, permitindo que as aplicações sejam livres de contexto. O processo foi dividido em quatro etapas (Selecionar *datasets*, Identificar conceitos, Recuperar instâncias e Alinhar dados). A etapa de alinhamento, por sua vez é composta de duas atividades (Alinhamento Simples e Alinhamento em Cascata).

A atividade de alinhamento em cascata define um subprocesso composto por quatro atividades (Recuperar instâncias de conceito relacionado, Alinhar instâncias de conceito relacionado, Recuperar instâncias de conceito principal e Alinhar instâncias do conceito principal). De modo a apoiar a realização das subatividades, nós apresentamos uma breve descrição para cada atividade.

No próximo capítulo será apresentada uma arquitetura que implementa o processo proposto.

Implementação do Processo

A execução do processo ocorre por meio de componentes para a análise de similaridade, persistência dos dados, geração do alinhamento, lógica entre as etapas e pré-processamento (ver Figura 13). Alguns componentes foram reusados (em preto), como o de similaridade léxica, que contém vários algoritmos para detectar a similaridade entre textos. Outros componentes foram desenvolvidos (em branco), como os responsáveis pela detecção da similaridade de recurso, alinhamento e outros.

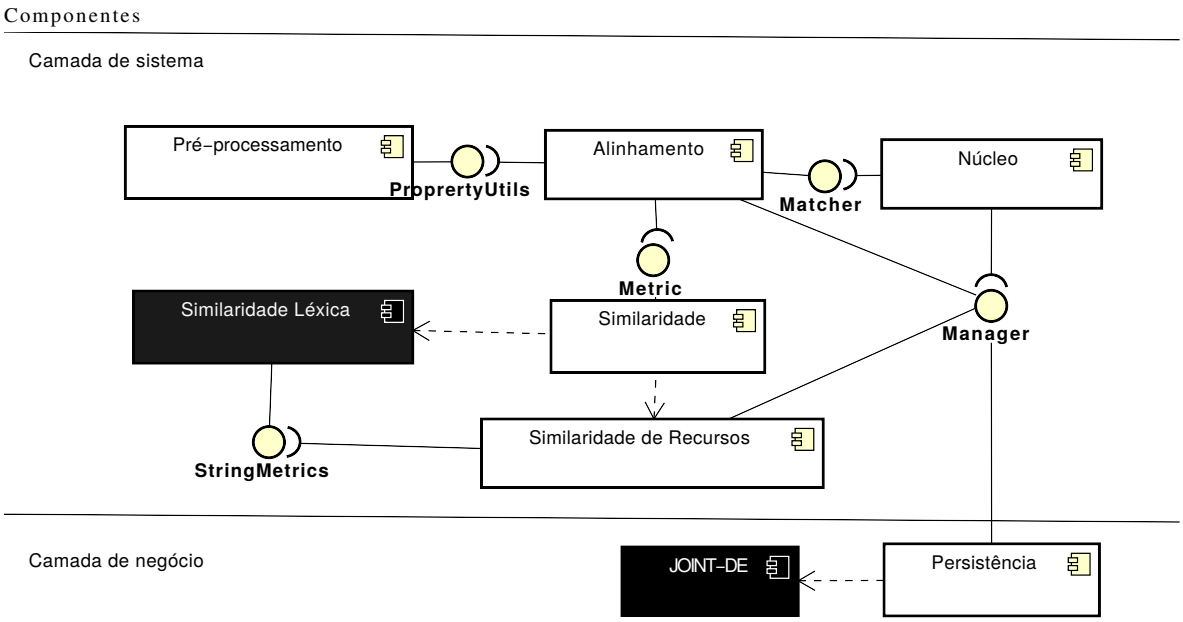


Figura 13 – Componentes utilizados na implementação do processo

Apesar de existirem diversas soluções que contém algoritmos para o cálculo de similaridade e alinhamento entre recursos, optou-se pelo desenvolvimento de uma abordagem que contemplasse problemas provenientes de bases de dados reais (acentuação, ausência de propriedades, formatação e outros) (CASTANO et al., 2011; FERRARA et al., 2008). Os principais componentes utilizados são:

Pré-processamento

O componente de pré-processamento tem como função realizar tratamentos nos textos que serão aplicados à função de similaridade. Alguns exemplos de tratamento são: tratamento de acentos, pontuação e outros.

Similaridade

O componente de similaridade é dividido em dois subcomponentes, sendo eles o de similaridade léxica e recurso. O primeiro utiliza métricas que analisam a similaridade entre palavras e textos. Algumas dessas técnicas são Levenshtein (LEVENSHTEIN, 1966), Cosseno (SINGHAL, 2001), Jaro-Winkler (WINKLER, 1990) e outros. O segundo componente, que se refere à similaridade de recurso utiliza o primeiro e tem como função gerar a similaridade entre recursos.

Para o cálculo da similaridade dos recursos foi utilizada uma abordagem baseada na técnica de subgrafo semântico (WANG; XU, 2008). Na prática, um subgrafo semântico diz respeito às triplas que estão relacionadas a um recurso qualquer, que estão de acordo com a modelagem ontológica. A Figura 9 representa um subgrafo que relaciona um autor e sua publicação. Além de propriedades de objetos, um subgrafo também possui propriedades de dados, tanto do recurso principal (Author), quanto dos conceitos relacionados (Production).

O valor gerado pelo componente de similaridade é enviado para o componente de alinhamento.

Alinhamento

O componente de alinhamento tem como responsabilidade determinar, de acordo com os valores obtidos na etapa de similaridade, se os recursos analisados realmente dizem respeito à mesma entidade do mundo real. Para determinar se o alinhamento deve ser realizado, este componente faz uso de limiares de aceitação, que são determinados previamente. Por esse motivo, o processo de alinhamento não é uma tarefa automática, pois precisa que os valores sejam ajustados. Existem diversas formas de determinar o valor do limiar, que vão desde executar várias vezes e analisar o melhor custo/benefício entre precisão e revocação até utilizar técnicas que atualizam o valor do limiar dinamicamente.

Núcleo

O componente de núcleo é responsável por concentrar e coordenar as configurações durante a execução. Atualmente, o componente de núcleo conta com três modalidades para a correspondência entre instâncias, sendo elas o alinhamento simples, que é executado em todas as modalidades, podendo também ser executado de forma independente, como mencionado na subseção 3.1.4.1. O alinhamento em cascata, que é executado quando é escolhido um conceito relacionado ao conceito cujas instâncias se deseja alinhar. O processo de execução do alinhamento em cascata pode ser encontrado na subseção 3.1.4.2. Por fim, há o alinhamento em multicascata, que ocorre quando mais de um conceito relacionado é escolhido. Vale ressaltar que esta variação

se trata dos alinhamentos em cascata de vários conceitos relacionados, que podem ser executados em paralelo.

Persistência

O componente de persistência é responsável por materializar as correspondências encontradas pelo componente de alinhamento. Para isso, é utilizado o JOINT (ver Figura 14), que de acordo com Holanda et al. (2013), é um arcabouço para facilitar o desenvolvimento de aplicações baseadas em ontologia. As *features* apresentadas pela ferramenta JOINT permitem que operações sejam realizadas diretamente no servidor de triplas (Virtuoso, OWLim etc.). Além disso, essa ferramenta também suporta a execução de consultas SPARQL, que através de um sistema de tradução, transforma as triplas em objetos da linguagem Java.

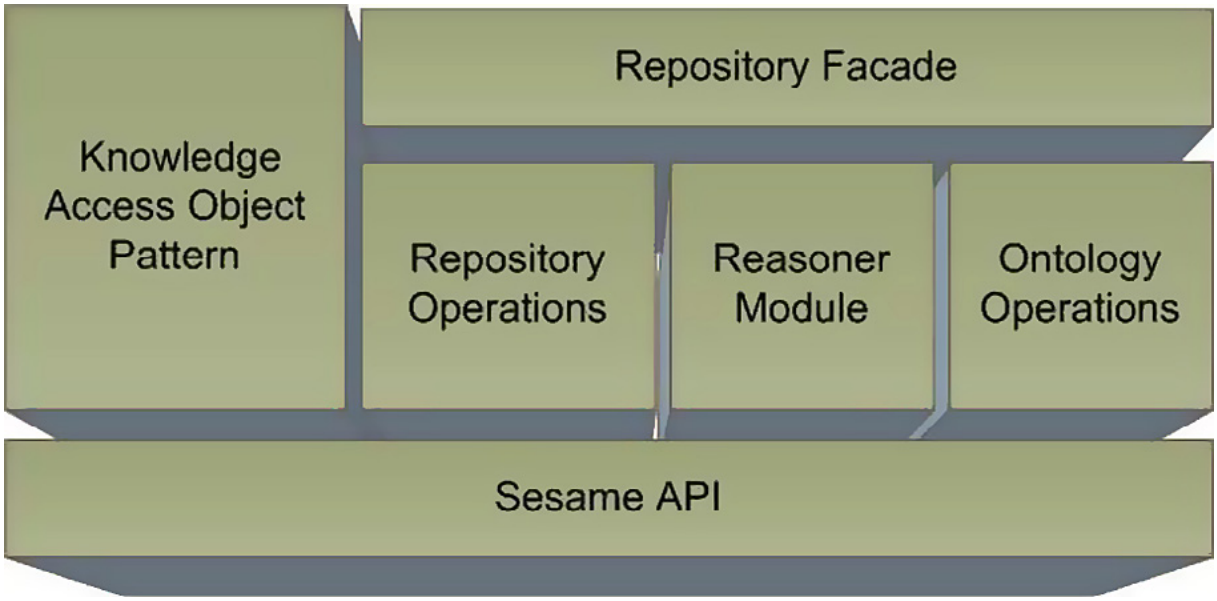


Figura 14 – Arquitetura da ferramenta JOINT
Fonte: (HOLANDA et al., 2013).

4 EXPERIMENTO

O presente trabalho apresentou um processo de alinhamento de dados independente de contexto. Esse processo foi consolidado em uma ferramenta para realizar a correspondência de instâncias. Este capítulo apresenta o experimento realizado para análise da eficácia do alinhamento de dados à partir do processo e ferramenta propostos. O capítulo está dividido em quatro seções que abordam (i) *Design* de experimento; (ii) Execução do experimento; (iii) Análise dos resultados e (iv) Principais conclusões obtidas.

Configuração de Experimento

Nesta seção, será detalhado o planejamento do experimento que foi projetado para este trabalho. Dentro do planejamento, encontra-se a definição da questão de pesquisa e derivação de hipóteses, a seleção das variáveis dependentes e independentes, a identificação da unidade experimental e a seleção do modelo experimental que será utilizado.

Situando o problema

A avaliação de ferramentas de correspondência de instâncias utiliza dois *datasets* e um alinhamento de referência. Como mencionado na subseção 1.3, isso permite que os desenvolvedores utilizem computação específica para os *datasets*. Com isso, retomamos o nosso problema:

(Problema Geral) Como identificar que duas instâncias referem-se à mesma entidade do mundo real? Atualmente, estratégias baseadas em similaridade, aprendizagem, regras e contexto (CASTANO et al., 2011) vêm sendo utilizadas para resolver o problema. Porém, segundo Homoceanu, Kalo e Balke (2014) essa ferramentas para correspondência de instância não estão prontas para alinhar dados automaticamente de forma confiável.

Diante disso, nos deparamos com nosso problema específico:

(Problema Específico) Como melhorar a eficácia das ferramentas para correspondência de instâncias?

Objetivos da Investigação

A pesquisa a ser realizada é de caráter experimental e tem como objetivo geral **avaliar a eficácia das ferramentas de correspondência de instâncias**. O intuito do experimento é **refutar** as hipóteses nulas definidas na subseção 4.1.3, indicando que a abordagem proposta, apesar de não conter computação específica para os *datasets*, é capaz de criar correspondências com eficácia.

Formalmente, o objetivo da nossa investigação pode ser definido como analisar ferramentas de correspondência de instâncias com a intenção de compará-las a respeito de sua eficácia no ponto de vista de geração de correspondência entre instâncias, no contexto de alinhamento de dados entre *datasets*, com o fim de utilizar as melhores abordagens provendo uma melhoria na qualidade das ferramentas.

Como objetivos específicos, temos:

i) Comparar a abordagem proposta e as existentes;

ii) Avaliar empiricamente a qualidade dos modelo criado.

Questões de Pesquisa e Hipóteses

Após a apresentação dos objetivos deste experimento, nos deparamos com a seguinte questão de pesquisa e suas respectivas hipóteses:

Q1 Como se comportam as ferramentas para correspondência de instâncias (AML, RiMOM-2016 e Proposta) com relação a eficácia?

A questão de pesquisa acima implica nas seguintes hipóteses de pesquisa:

- H1-0: A precisão apresentada pelas abordagens é igual.
- H1-1: A precisão apresentada pelas abordagens é diferente.
- H2-0: A revocação apresentada pelas abordagens é igual.
- H2-1: A revocação apresentada pelas abordagens é diferente.
- H3-0: A medida-f apresentada pelas abordagens é igual.
- H3-1: A medida-f apresentada pelas abordagens é diferente.

Fatores e Variáveis de Resposta

A partir da definição das hipóteses na subseção 4.1.3, temos os fatores, também conhecidos como variáveis independentes, como sendo:

- **Ferramenta:** Esta variável especifica qual a ferramenta de alinhamento será avaliada

Como variáveis de resposta, também conhecidas como variáveis dependentes, nós temos:

Precisão (P): Para a correspondência de instâncias, indica a quantidade de correspondências que são relevantes em relação a todas as correspondências geradas pelas ferramentas. A definição formal para esta métrica é descrita pela equação abaixo:

$$P = \frac{|A \cap B|}{|A|} \tag{4.1}$$

Revocação (R): Para a correspondência de instâncias, indica a quantidade de correspondências relevantes com relação ao conjunto de todas as correspondências possíveis (espelho). A definição formal para esta métrica é descrita pela equação abaixo:

$$R = \frac{|A \cap B|}{|B|} \tag{4.2}$$

Medida-f (F): Média harmônica entre precisão e revocação. A intenção é transformar essas duas métricas em apenas uma. A definição formal para esta métrica é descrita pela equação abaixo:

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{4.3}$$

Níveis dos Fatores

Os níveis dos fatores são apresentados na Tabela 3.

Tabela 3 – Níveis de fatores

Fator	Nível	Descrição
Ferramenta	F1	Ferramenta AML
	F2	Ferramenta RiMOM-2016
	F3	Proposta

Definição formal das Hipóteses

Formalmente, todas as hipóteses definidas na seção 4.1.3 podem ser definidas conforme a Tabela 4. P, R e F são funções que retornam, respectivamente a precisão, revocação e medida-f, com relação às abordagens F1 (AML), F2 (RiMOM-2016) e F3 (Proposta).

Tabela 4 – Formalização das hipóteses

Hipótese	Hipótese Nula	Hipótese Alternativa
H1	H1-0:$P(F_i) = P(F_j); i, j \in \{1, 2, 3\}; i \neq j$	H1-1:$\exists i, j \in \{1, 2, 3\}; i \neq j; P(F_i) \neq P(F_j)$
H2	H2-0:$R(F_i) = R(F_j); i, j \in \{1, 2, 3\}; i \neq j$	H2-1:$\exists i, j \in \{1, 2, 3\}; i \neq j; R(F_i) \neq R(F_j)$
H3	H3-0:$F(F_i) = F(F_j); i, j \in \{1, 2, 3\}; i \neq j$	H3-1:$\exists i, j \in \{1, 2, 3\}; i \neq j; F(F_i) \neq F(F_j)$

Unidades Experimentais

Levando em consideração as diversas classificações de experimento (MONTGOMERY, 2012), o presente experimento é classificado como fatorial completo com blocagem. A blocagem foi escolhida com o objetivo de suprimir os efeitos dos *datasets* nas variáveis resposta. Para cada cenário houve a execução de todos os níveis de fatores, garantindo a completude do experimento.

Plano de execução

A Figura 15 apresenta os passos de execução de cada alinhamento, que são descritos abaixo:

- Construir contêiner com as configurações zeradas;
- Carregar dos dados;
- Executar ferramenta dentro do contêiner;
- Coletar dados de alinhamento;
- Destruir contêiner;
- Analisar dados.

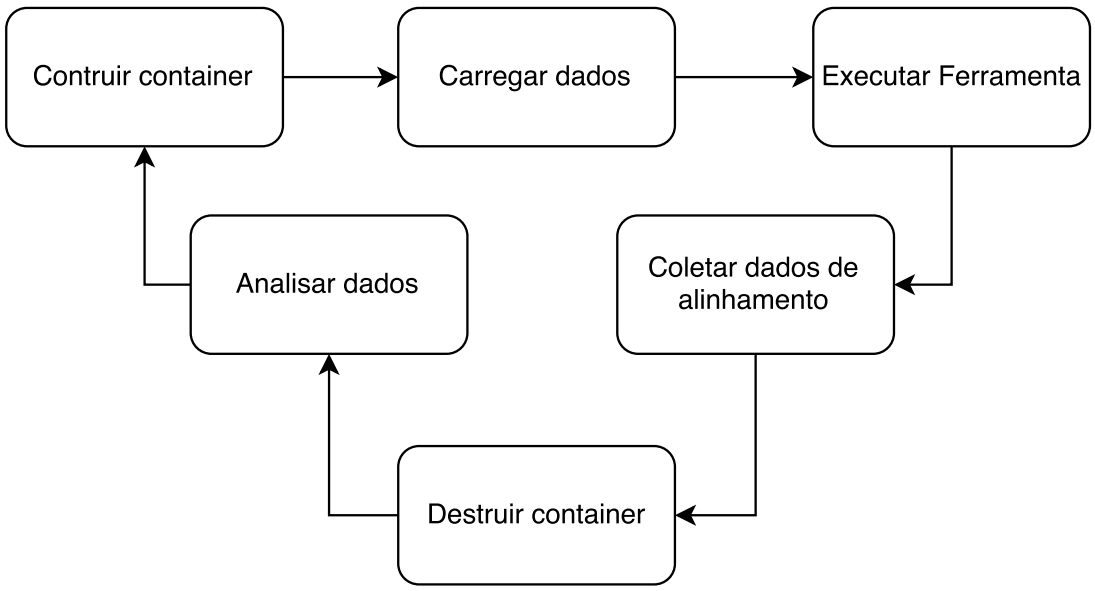


Figura 15 – Processo de execução do experimento

Coleta dos Dados

Os dados utilizados neste estudo foram fornecidos pela OAEI. O *dataset*[1] é composto por três variações (*9-heterogeneities*, *4-heterogeneities* e *falsepositives-trap*). Cada uma das variações conta com um alinhamento de referência, para que seja possível calcular as métricas.

Execução do experimento

Para avaliar as ferramentas para correspondência de instância, o experimento irá contar com dois cenários possíveis, com uma execução para cada nível dos fatores definidos na subseção 4.1.5, totalizando 6 execuções. A Tabela 5 apresenta os cenários.

Tabela 5 – Cenários de alinhamento

Cenário	Datasets
C1	9-heterogeneities
C2	falsepositives-trap

Instrumentação

Para a realização do experimento de correspondência de instâncias, serão utilizados os seguintes instrumentos.

- Intellij IDEA 2016.3 para desenvolvimento do código e execução da proposta.
- Virtuoso RDF Store - 07.20.3217;
- OpenJDK 64-Bit Server VM (build 25.111-b14, mixed mode)

Para isolar os efeitos entre as execuções, todo o experimento foi conduzido em um contêiner, que permite isolar as aplicações, fazendo com que as aplicações sejam executadas em ambientes idênticos sem que gerem efeitos colaterais entre si.

Ameaças à validade

Embora todo experimento tenha sido projetado para minimizar possíveis ameaças que comprometam suas conclusões, existem algumas ameaças que devem ser mencionadas. Nos tópicos a seguir, serão apresentadas e detalhadas as ameaças deste experimento.

[1] <http://islab.di.unimi.it/content/im_oaei/2016/#doremus>

- **Ameaças à validade interna:**
  Uma possível ameaça interna à validade do experimento pode ser a seleção das unidades experimentais, pois os *dataset* utilizados no experimento foram fornecidos pelo OAEI e nenhum outro *dataset* foi utilizado.

- **Ameaças à validade externa**
  As unidades experimentais desta pesquisa são selecionadas de apenas uma base de dados correspondente a cada modelo, tendo, cada uma delas, características que poderão não ser generalizadas para as demais bases

- **Ameaças à validade de construto**
  É possível que a quantidade de fatores e cenários não sejam suficientes para observar diferenças significativas na eficácia entre as abordagens utilizadas para a correspondência de instâncias. Além disso, deve-se considerar que o tempo de resposta não foi considerado no experimento.

- **Ameaças à validade de conclusão**
  Devido à pequena quantidade de dados por *dataset*, é possível que a quantidade de instâncias por *dataset*, não seja suficiente para observar diferenças significativas nas métricas associadas.

Análise dos Resultados

O experimento foi conduzido de acordo com o planejamento descrito anteriormente neste capítulo. Depois da execução, os *datasets* de alinhamento foram coletados. Os dados de precisão, revocação e medida-f foram calculados de acordo com as funções especificadas anteriormente. Ao longo desta seção, uma análise descritiva dos dados obtidos será apresentada com as métricas referentes a cada cenário.

As figuras 16 e 17 apresentam as ferramentas de acordo com seus resultados em cada um dos cenários. A Tabela 6 sumariza os dados obtidos para cada cenário.

Tabela 6 – Sumarização dos dados relativos a precisão, revocação e medida-f por cenário.

Cenário	Ferramenta	Precisão	Revocação	Medida-F
C1	Proposta	1	0.875	0.933
	AML	0.966	0.875	0.918
	RiMOM	0.813	0.813	0.813
C2	AML	0.921	0.854	0.886
	Proposta	0.906	0.707	0.794
	RiMOM	0.707	0.707	0.707

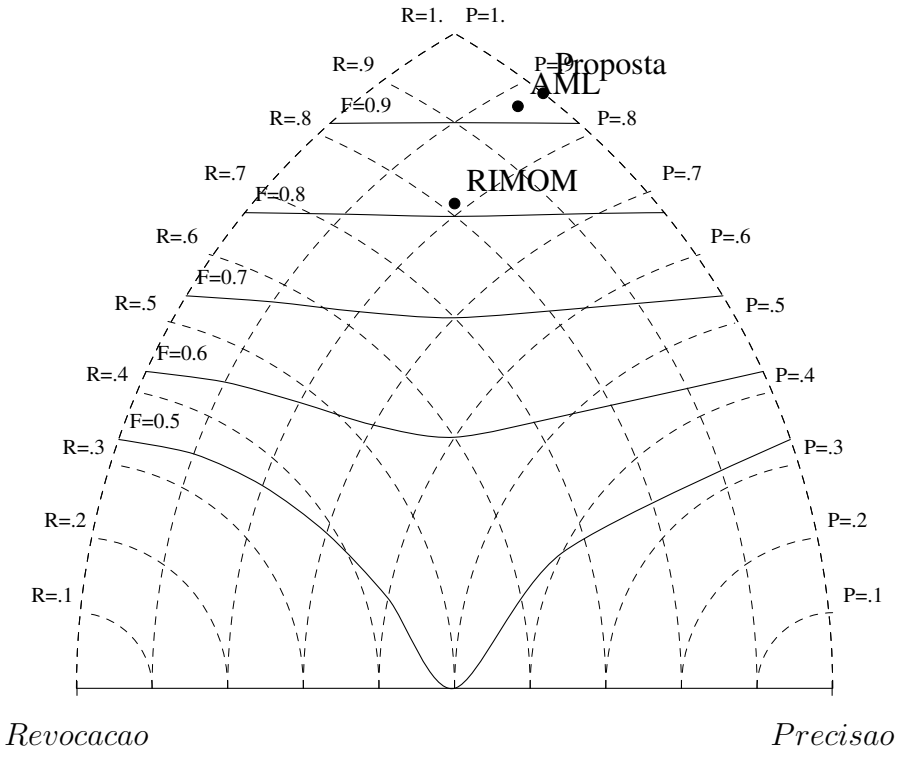


Figura 16 – Resultado das ferramentas para o cenário 1

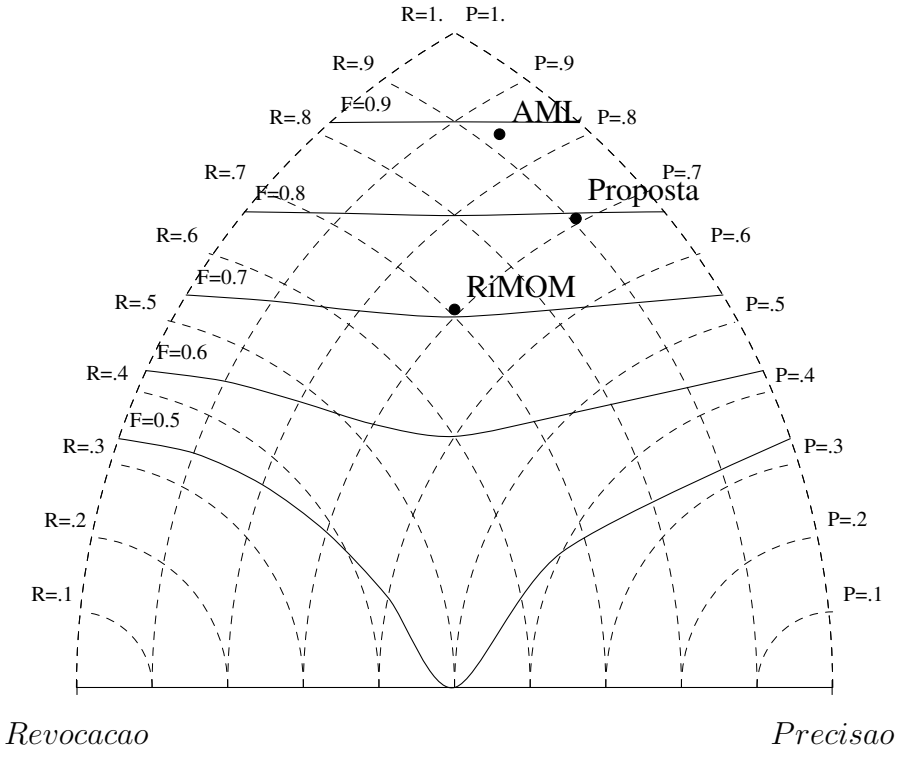


Figura 17 – Resultado das ferramentas para o cenário 2

A seguir serão apresentados os resultados mapeados para suas respectivas hipóteses. A primeira hipótese refere-se à precisão apresentada pelas ferramentas. As hipóteses nula e alternativa estão numeradas abaixo, respectivamente:

- **H1-0:** A precisão apresentada pelas abordagens é igual.
- **H1-1:** A precisão apresentada pelas abordagens é diferente.

Para realizar a análise, devemos observar a precisão nos dois cenários de alinhamento. Como descrito na Tabela 6, a proposta apresentou precisão de 1 para o cenário 1 e 0.906 para o cenário 2.

A próxima validação de hipótese se refere à revocação. As hipóteses nula e alternativa estão numeradas abaixo, respectivamente:

- **H2-0:** A revocação apresentada pelas abordagens é igual.
- **H2-1:** A revocação apresentada pelas abordagens é diferente.

Diante do resultado apresentado pela proposta nos dois cenários de alinhamento, onde apresentou uma revocação de 0.875 para o cenário C1 e 0.707 para o cenário C2. Apesar do bom desempenho, a proposta apresentou valores iguais a pelo menos uma das ferramentas em cada um dos cenários.

Finalizando a verificação de hipóteses, são analisadas as hipóteses relacionadas à medida-f. As hipóteses nula e alternativa estão listadas abaixo, respectivamente:

- **H3-0:** A medida-f apresentada pelas abordagens é igual.
- **H3-1:** A medida-f apresentada pelas abordagens é diferente.

Conforme os resultados apresentados na Tabela 6, a proposta ficou em primeira posição no cenário 1 com uma medida-f de 0.933 e segunda posição no cenário 2 com 0.794.

Para a análise estatística inferencial, visando a validação das hipóteses, utilizamos o teste de Fisher (FISHER, 1922) para comparar os pares das métricas. Esse teste pode ser usado para analisar a significância estatística da amostra. Com isso, é possível aceitar ou refutar qualquer uma das hipóteses, entre outras palavras, verificar se os resultados obtidos possuem diferenças estatísticas. Para isso, usou-se os resultados obtidos no experimento como entrada, seguindo a configuração apresentada na Tabela 6. Como resultado, o teste estatístico apresentou p-valor igual a 0.8333, indicando que as ferramentas apresentam eficácia similar em relação às métricas. Consequentemente, não é possível refutar nenhuma das hipóteses nulas.

Principais conclusões

O experimento conduzido neste capítulo, visou avaliar a eficácia das ferramentas de alinhamento de dados conectados com relação às métricas de precisão, revocação e medida-f. Essas variáveis foram avaliadas separadamente em cada um dos dois cenários C1 e C2. No cenário C1, os dados apresentam 9 tipos de heterogeneidade (*e.g.*, multilinguagem, diferença nos catálogos, diferença fonética, diferentes graus de descrição). No cenário C2, os dados apresentam conjuntos instâncias similares, havendo apenas uma correspondência possível, o que caracteriza as demais instâncias como falso-positivas.

Diante dos resultados apresentados pelo teste estatístico, tem-se que as ferramentas apresentam eficácia similar para ambos os cenários analisados. Contudo, pelas análises realizadas sobre as métricas, foi possível verificar que a proposta se destacou apenas em um dos cenários, o C1. Entretanto, deve-se ressaltar que a bordagem proposta não utiliza implementações específicas para os *datasets* analisados, permitindo que seja utilizada em outros contextos com facilidade.

5 ESTUDO DE CASO (QIE)

Neste capítulo será descrita a utilização do processo de alinhamento através de um estudo de caso, o QIE. O QIE foi escolhido, pois se trata de um projeto que pretende cruzar informações sobre a comunidade brasileira de Informática na Educação. Este capítulo está estruturado em três seções. A seção 5.1 apresenta uma descrição do estudo de caso e a conversão dos dados para RDF. A seção 5.2 apresenta o passo a passo de como o processo foi executado no estudo de caso. Por fim, a seção 5.3 apresenta os resultados apresentados a partir do alinhamento.

Descrição do estudo

Para validar a eficácia da solução, foi solicitado a membros da Comissão Especial de Informática na Educação que sugerissem perguntas de interesse sobre a comunidade brasileira de Informática na Educação. Como resultado foi levantado um conjunto contendo mais de 30 questões. A Tabela 7 apresenta as questões propostas pelos membros.

De acordo com as perguntas realizadas, pode-se perceber que para responder algumas delas é necessário o cruzamento de diferentes fontes de informações dos pesquisadores e suas publicações, sendo elas a Revista Brasileira de Informática na Educação[1] (RBIE), Workshop de Informática na Escola[2] (WIE), Simpósio Brasileiro de Informática na Educação[3] (SBIE) e curriculum Lattes[4]. Vale a pena ressaltar que os *datasets* foram disponibilizados como arquivos XML, sendo necessário transformá-los para RDF.

Para modelar os dados, foram utilizadas as ontologias dac[5] e lattes[6]. A primeira tem o objetivo de modelar o domínio de publicação (ver Figura 18). A segunda foi construída para modelar o domínio do lattes (ver Figura 19).

Para transformar os dados para RDF foi utilizada a ferramenta OpenRefine[7] com a extensão para suportar RDF. Essa ferramenta foi selecionada devido a sua facilidade para criar os *templates* de transformação. Após a transformação dos dados, as ontologias e os dados foram persistidos no Virtuoso[8]. A Figura 20 representa o processo de conversão dos dados.

Com o processo de conversão dos dados foram geradas 1,1 milhão de triplas, sendo distribuídas da seguinte forma: (96%, 1.094.307 triplas) pertencem ao Lattes, (1,61%; 18.363 triplas) pertencem ao SBIE, (1,21%; 14.601 triplas) pertencem ao WIE e (1,1%; 12.503 triplas)

1 <http://www.br-ie.org/pub/index.php/rbie>
2 <http://www.br-ie.org/pub/index.php/wie>
3 <http://www.br-ie.org/pub/index.php/sbie>
4 <http://lattes.cnpq.br>
5 <https://github.com/josmarios/dac/blob/master/Ontologies/dacV2.1.owl>
6 <https://github.com/armandobs14/lattes/blob/master/lattes.owl>
7 <http://openrefine.org>
8 <https://virtuoso.openlinksw.com>

Tabela 7 – Perguntas sugeridas pela comunidade

ID	Questões
Q01	Quantos pesquisadores de Informática na Educação (IE) há na comunidade?
Q02	Quais são os pesquisadores em IE no Brasil?
Q03	Onde estão os pesquisadores de IE no Brasil? (Estado)
Q04	Onde estão trabalhando os pesquisadores de IE no Brasil? (Universidade)
Q05	Quais pesquisadores de IE no Brasil são doutores?
Q06	Quantos pesquisadores de IE no Brasil são doutores?
Q07	Quem dos pesquisadores de IE no Brasil possui marca registrada?
Q08	Onde os pesquisadores de IE no Brasil fizeram o Doutorado?
Q09	Onde os pesquisadores de IE no Brasil fizeram os pós-doutorados?
Q10	Quantas publicações o autor “z” tem no evento “x” (SBIE/WIE) da área de IE?
Q11	Quantos trabalhos foram publicados no evento “x” (SBIE/WIE) da área de IE?
Q12	Quantos autores publicaram no evento “x” (SBIE/WIE) da área de IE?
Q13	Quantos artigos foram publicados no periódico “y” (RBIE) da área de IE?
Q14	Lista de doutores que publicaram na RBIE e seus e-mails.
Q15	Lista de autores da comunidade de IE no Brasil, com competência e e-mail.
Q16	Lista de artigos publicados na RBIE - Geral.
Q17	Quantos pesquisadores da comunidade de IE são bolsistas PQ/DT e qual o nível?
Q18	Quais são as principais competências da comunidade de IE?
Q19	Quais conceitos são explorados pelos pesquisadores de IE no Brasil?
Q20	Quais os temas são mais pesquisados em IE no Brasil?
Q21	Quais pesquisadores de IE no Brasil colaboram entre si?
Q22	Quais instituições que os pesquisadores de IE no Brasil atuam que colaboram entre si?
Q23	Quais são os trabalhos relacionados publicados no SBIE, WIE e RBIE?
Q24	Como os conceitos explorados nas publicações de IE evoluem ao longo do tempo?
Q25	Mapa de tendências de pesquisa em IE no Brasil em uma linha do tempo.
Q26	O quão um pesquisador X está publicando no SBIE, WIE e RBIE ao longo do tempo?
Q27	Lista de bolsistas de produtividade de pesquisadores em IE no Brasil.
Q28	Quais as instituições que os pesquisadores de IE no Brasil atuam?
Q29	Quais autores da comunidade de IE no Brasil publicaram na conferência "X"?
Q30	Quantos pesquisadores de IE no Brasil estão em Programas de Pós-Graduação de Computação?
Q31	Quem são os maiores especialistas em recursos digitais e objetos de aprendizagem no Brasil?

pertencem ao RBIE.

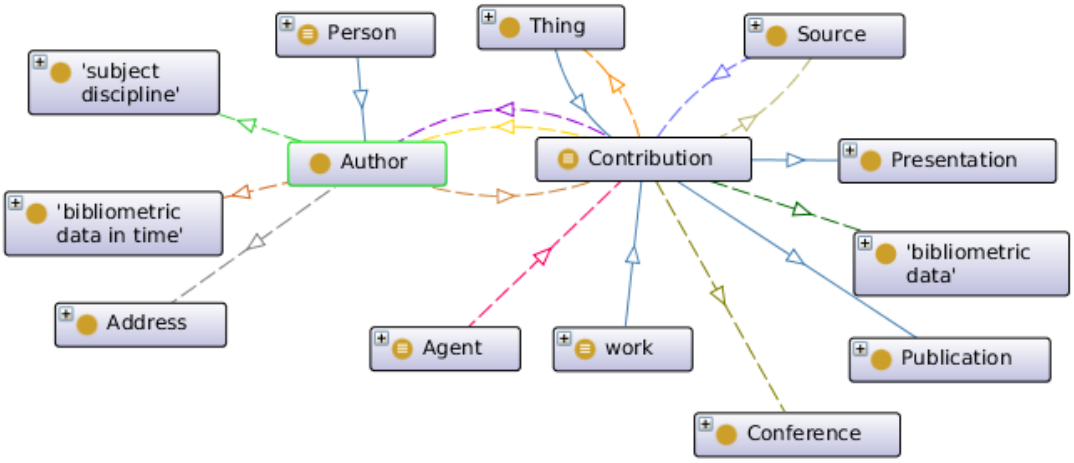


Figura 18 – Taxonomia da ontologia dac

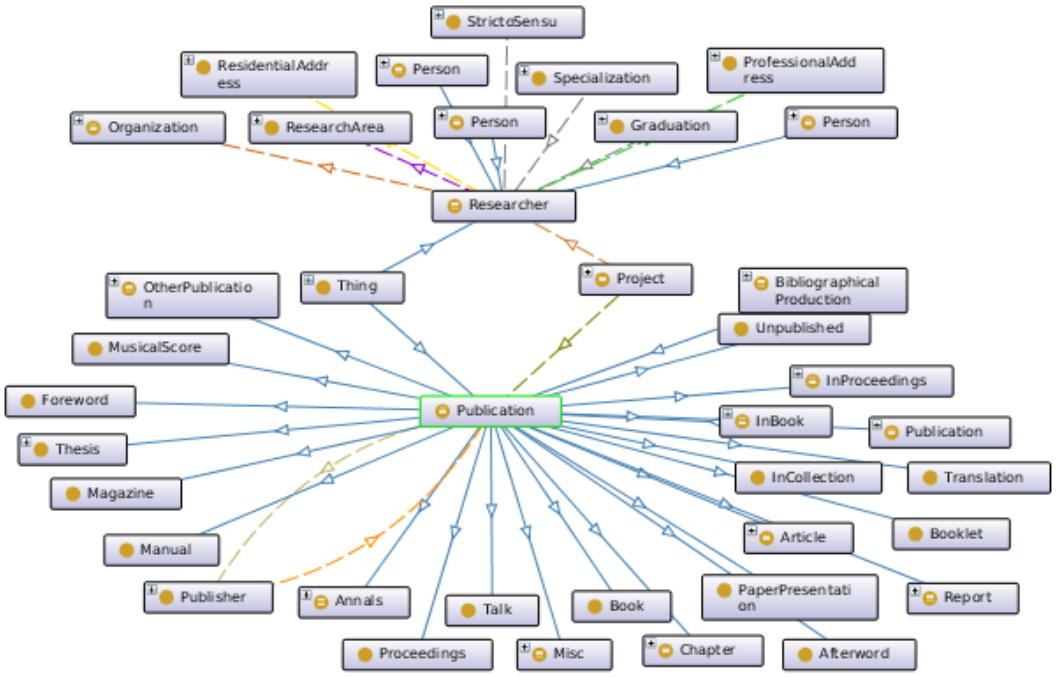


Figura 19 – Taxonomia da ontologia lattes

Execução do processo

Nesta seção descreve-se como foi realizada cada uma das etapas do processo de correspondência.

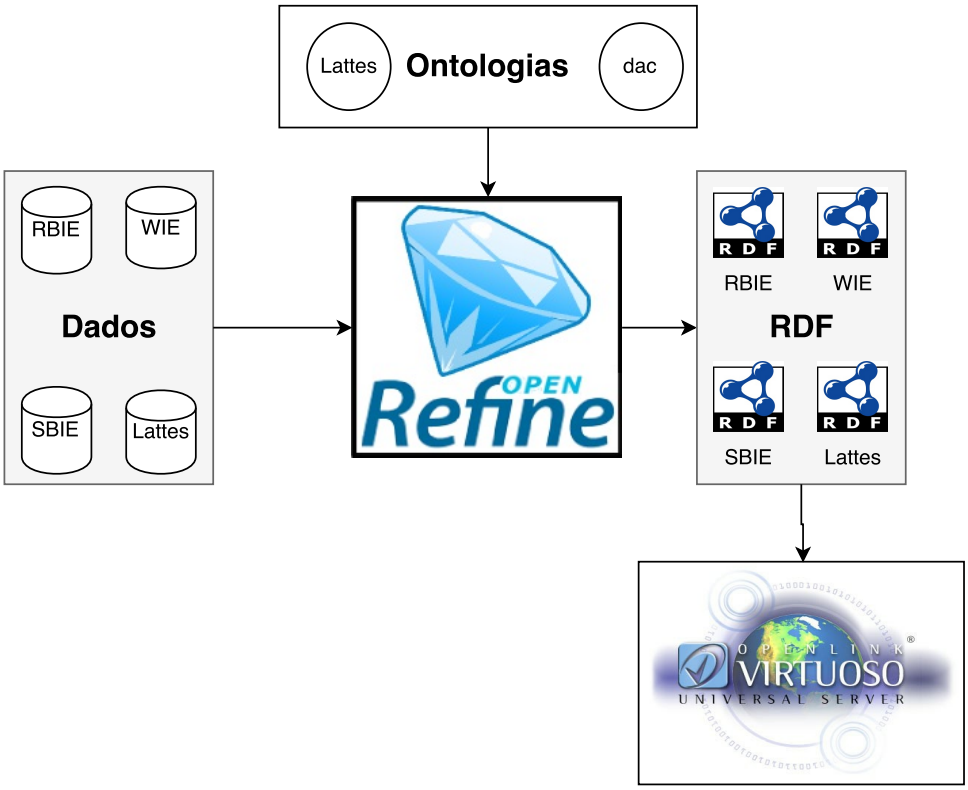


Figura 20 – Processo de conversão para rdf

Selecionar *Datasets*

Esta etapa refere-se à seleção de quais *datasets* são utilizados como entrada para o processo de correspondência de instâncias. Vale ressaltar que dois ou mais *datasets* podem ser selecionados. Neste contexto, foram selecionados os *datasets* do RBIE, SBIE, WIE e Lattes.

Identificar Conceitos

Esta etapa do processo expõe a seleção dos conceitos (principal e relacionados), que são utilizados no processo. Atualmente existe apenas uma restrição quanto a seleção dos conceitos, nesta é possível selecionar apenas um conceito principal.

Conceito Principal

Como descrito na seção 3.1.2, para facilitar a identificação do conceito principal por parte do usuário, a consulta 3.1 foi desenvolvida. A Tabela 8 apresenta os resultados obtidos através execução da consulta.

O conceito ***Author***, que representa a segunda maior quantidade de instâncias nos dados, foi selecionado como conceito principal. A escolha desse conceito deu-se não pela quantidade

Tabela 8 – Conceito da ontologia e quantidade de instâncias

Conceitos	quantidade
http://www.ic.ufal.br/dac/Contribution	868752
http://www.ic.ufal.br/dac/Author	155680
http://www.ic.ufal.br/lattes/DoctoralDegree	2387
http://www.ic.ufal.br/lattes/Graduation	2195
http://www.w3.org/2002/07/owl#Class	1680
http://www.ic.ufal.br/lattes/Course	1186
http://www.w3.org/1999/02/22-rdf-syntax-ns#List	204
http://www.w3.org/2002/07/owl#Restriction	161
http://www.w3.org/2002/07/owl#ObjectProperty	128
http://www.w3.org/2000/01/rdf-schema#Class	60
http://www.w3.org/2002/07/owl#Ontology	24
http://www.w3.org/1999/02/22-rdf-syntax-ns#Property	23
http://purl.org/dc/terms/AgentClass	3

de instâncias, mas por questões estratégicas, visto que a proposta deste estudo de caso é cruzar informações de pesquisadores.

Conceito Relacionado

Para selecionar o conceito relacionado que será utilizado, foi executada a consulta 3.2. A Tabela 9 apresenta os conceitos relacionados ao conceito principal previamente selecionado.

Tabela 9 – Lista com conceitos relacionados

Conceito Relacionado
http://www.ic.ufal.br/dac/Contribution
http://www.ic.ufal.br/lattes/Software
http://www.ic.ufal.br/lattes/TradeMark
http://xmlns.com/foaf/0.1/Organization
http://www.ic.ufal.br/lattes/Organization
http://www.nees.com.br/boa-moradia/crawler/LocalityAddress
http://www.ic.ufal.br/lattes/DoctoralDegree
http://www.ic.ufal.br/lattes/Graduation
http://www.ic.ufal.br/lattes/MastersDegree

O conceito ***Contribution*** foi selecionado como conceito relacionado. Esse conceito foi utilizado durante o alinhamento em cascata. Vale ressaltar que mais de um conceito relacionado pode ser selecionado.

Listar Recursos

A lista de recursos é gerada de forma automática com base nos conceitos selecionados anteriormente. A partir da lista de recursos são montados os pares candidatos. Vale ressaltar que no estudo em questão um *dataset* pode conter mais de uma instância para a mesma entidade do mundo real (e.g. mais de uma URI para o mesmo pesquisador). Dessa forma, foram gerados pares candidatos dentro do mesmo *dataset*, caracterizando o alinhamento interno.

Alinhar Dados

A etapa de alinhamento é responsável por determinar a correspondências entre as instâncias. Neste processo existem duas abordagens de alinhamento sendo elas simples e em cascata. Na abordagem simples, os recursos são comparados diretamente, explorando as propriedades e suas características. Na abordagem em cascata os recursos são comparados a partir dos recursos relacionados.

Alinhamento Simples

Assim como outras abordagens para a correspondência de instâncias (ZHANG et al., 2016), também foram utilizadas funções que analisam a similaridade entre dois recursos. Para determinar a similaridade entre os pares foi utilizada a função 3.2. Essa função de similaridade gera valores entre 0 e 1, sendo 0 totalmente distintos e 1 iguais. Além da função utilizada, foram definidos limiares (*threshold*). Isso quer dizer que valores de similaridade acima do limiar eram considerados correspondentes.

Inicialmente o valor do limiar foi definido de forma arbitrária e posteriormente ajustado com ajuda de testes. Para isso, o mesmo *dataset* era alinhado diversas vezes utilizando um valor de limiar para cada execução. Por fim, o valor do limiar foi estabelecido em 0.88.

Alinhamento em Cascata

O alinhamento em cascata consiste em alinhar instâncias do recurso principal a partir de recursos relacionados. Esta etapa do processo é executada para cada um dos conceitos relacionados selecionados na etapa de identificação de conceitos e consiste de três atividades:

- **Alinhar recursos relacionados:** O alinhamento simples é executado entre as instâncias que pertencem ao conceito relacionado;
- **Recuperar instâncias do conceito principal:** A partir do alinhamento entre instâncias de recurso relacionado, as instâncias que pertencem ao conceito principal são recuperadas.

- **Alinhar instâncias do conceito principal:** A partir das instâncias recuperadas novos pares candidatos são gerados e passados como entrada para o alinhamento simples.

Resultados

Os resultados apresentados nessa seção estão separados em duas partes. A primeira consiste dos alinhamentos gerados. A segunda aborda as respostas das perguntas realizadas pela comunidade.

Alinhamentos

Após a realização do alinhamento por parte da ferramenta, foi realizado um levantamento, a partir do qual foi possível gerar informações como a quantidade de recursos são repetidos nas bases, total de recursos alinhados com o perfil Lattes, bem como precisão, revocação e medida-f (GOUTTE; GAUSSIER, 2005) para analisar a confiabilidade dos alinhamentos (ver Tabela 10). Vale ressaltar que o alinhamento de referência foi gerado manualmente com o auxílio de especialistas no domínio.

Tabela 10 – Resultado dos alinhamentos com relação aos dados do Lattes

Dataset	**RBIE**	**SBIE**	**WIE**
Iniciais	1118	1687	1952
Finais	806	1032	1098
Perfis com Lattes*	92.03% (1029)	71.54% (1207)	75.20% (1468)
Perfis com Lattes**	89.06% (717)	70.16% (792)	67.48% (741)
P | R | F	0.97 | 1 | 0.98	0.94 | 1 | 0.97	0.84 | 1 | 0.91

* - com repetição ** - sem repetição

Respostas

O processo de correspondência de instância permite não só identificar as instâncias que se referem a mesma entidade, mas também permite que informações complementares sejam integradas. Dessa forma, foi necessário consultar mais de uma base ao mesmo tempo para responder às perguntas feitas pela comunidade. Devido à quantidade de perguntas feitas, apenas algumas delas serão apresentadas a seguir.

Outro fator que deve ser ressaltado trata-se de problemas nos dados que eram fornecidos pelos autores (quantidade de autores diferentes para o mesmo artigo, formatação do nome). Além disso, foi possível notar também o fornecimento de informação inverídica, onde 77 autores usaram o mesmo e-mail (autor@email.com).

Q03 - Onde estão os pesquisadores de IE no Brasil? (Estado)

Através desta pergunta é possível saber onde os pesquisadores que publicaram no RBIE, SBIE ou WIE trabalham. Essa informação é obtida através do endereço profissional que pode ser encontrado no perfil do curriculum Lattes de cada pesquisador. Desta forma, para responder a esta pergunta, é necessário identificar o perfil desses pesquisadores no curriculum Lattes.

A consulta 5.1 recupera o endereço profissional dos pesquisadores que publicaram no WIE. Na linha 9, a transitividade da propriedade *owl:sameAs* é utilizada com o objetivo de recuperar todos os perfis correspondentes. Para que a consulta não entrasse em loop, foi estabelecido a consulta de até 5 elementos compondo a transitividade. A quantidade de 5 elementos foi escolhida de forma manual, com esse valor foi possível alcançar todas as propriedades possíveis através da transividade.

```
1  PREFIX dac:<http://www.ic.ufal.br/dac/>
2  PREFIX lattes:<http://www.ic.ufal.br/lattes/>
3  SELECT  ?uf count (distinct ?o) as ?total
4  FROM <http://www.ic.ufal.br/dac/wie/>
5  FROM <http://www.ic.ufal.br/dac/author/wie/lattes/alignments/>
6  FROM <http://www.ic.ufal.br/dac/author/wie/wie/alignments/>
7  FROM <http://www.ic.ufal.br/dac/lattes/>
8  WHERE {
9  ?s a dac:Author; owl:sameAs{,5} ?o.
10 filter regex(?s,"http://www.ic.ufal.br/dac/author/wie/(\\d)+$
      ","i")
11 filter regex(?o,"http://www.ic.ufal.br/dac/author/lattes/(.)+$
      ","i")
12 filter not exists {graph<http://www.ic.ufal.br/dac/author/wie/
      wie/alignments/>{?k owl:sameAs ?s}}
13 ?o lattes:hasProfessionalAddress ?address.
14 ?address lattes:uf ?uf
15 filter exists {?elem owl:sameAs ?g}
16 }
17 group by ?uf
18 order by desc(?total)
```

Código 5.1 – Consulta para recuperar a concentração de pesquisadores por UF

A Figura 21 apresenta a concentração de pesquisadores por unidade federativa (UF). Este mapa também conta com uma versão iterativa[9].

[9] <https://fiddle.jshell.net/4hernfu6/3/show/>

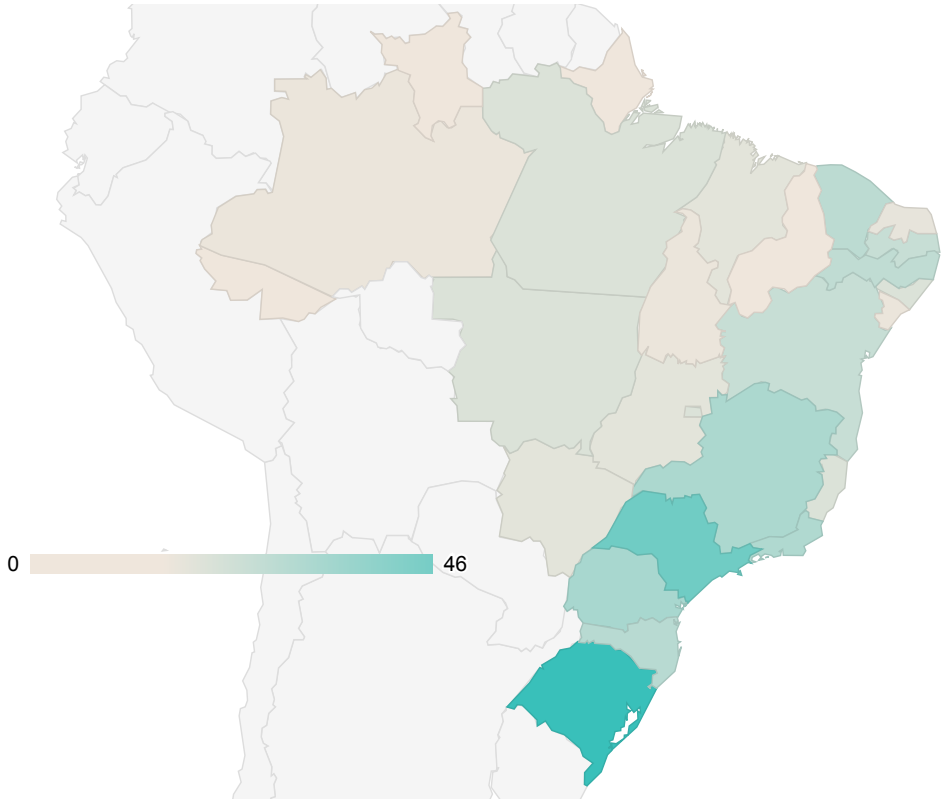


Figura 21 – Concentração de pesquisadores por UF

Q08 - Onde os pesquisadores de IE no Brasil fizeram o Doutorado?

Através desta pergunta é possível saber onde os pesquisadores que publicaram no RBIE, SBIE e WIE concluíram seus doutorados. Assim como a informação de endereço profissional, esta informação também pode ser obtida através do cruzamento entre essas bases e o Lattes. A consulta 5.2 recupera a instituição onde os pesquisadores que publicaram no WIE concluíram seus doutorados.

```
1
2  PREFIX dac:<http://www.ic.ufal.br/dac/>
3  PREFIX lattes:<http://www.ic.ufal.br/lattes/>
4
5  Select ?nameInstitution count( distinct ?g) as ?Count
6  FROM <http://www.ic.ufal.br/dac/wie/>
7  from <http://www.ic.ufal.br/dac/author/wie/lattes/alignments/>
8  from <http://www.ic.ufal.br/dac/author/wie/wie/alignments/>
9  FROM <http://www.ic.ufal.br/dac/lattes/>
10
11 where {
12 ?s a dac:Author; owl:sameAs* ?g.
13 ?g foaf:name ?oname.
```

```
14 filter regex(?s,"http://www.ic.ufal.br/dac/author/wie/(\\d)+$
      ","i")
15 filter regex(?g,"http://www.ic.ufal.br/dac/author/lattes/(.)+$
      ","i")
16 filter not exists {graph<http://www.ic.ufal.br/dac/author/wie/
      wie/alignments/>{?k owl:sameAs ?s}}
17 ?g lattes:hasAcademicDegree ?t.
18 ?t a lattes:DoctoralDegree.
19 ?t lattes:hasInstitution ?institution.
20 ?institution foaf:name ?nameInstitution
21 }
22 Order by desc(?Count)
```

Código 5.2 – Consulta para recuperar a concentração de doutorados por universidade

A Figura 22 apresenta a concentração de doutorados concluídos por universidade.

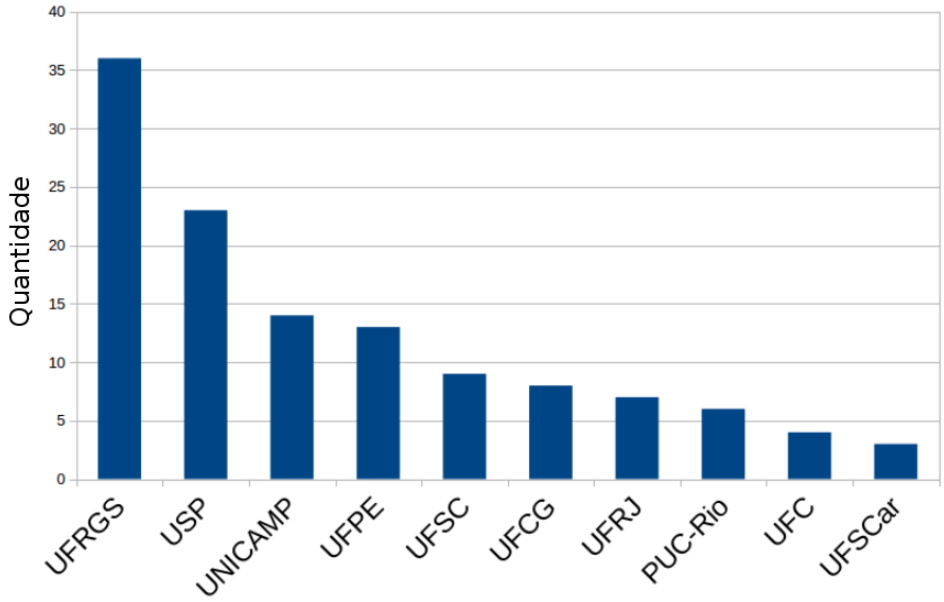


Figura 22 – Quantidade pesquisadores com doutorado por universidade

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O trabalho apresentou uma abordagem semiautomática para alinhar *datasets* reais. Através de diagramas de atividades e de componentes foi descrito o processo utilizado para conectar recursos de diferentes *datasets*. Essa proposta se faz necessária justamente pela necessidade de soluções capazes de alinhar dados de forma confiável e com menor conhecimento possível do domínio. Além disso, a solução permite que o alinhamento seja executado diretamente dentro do armazenamento de triplas, não havendo a necessidade de gerar arquivos para alinhar.

Visando avaliar a abordagem proposta em um cenário real, um estudo de caso foi desenvolvido. Neste estudo, a proposta foi utilizada para alinhar *datasets* do RBIE, SBIE, WIE e Lattes. A utilização da solução de alinhamento permitiu que diversas perguntas pudessem ser respondidas. Além disso, foi possível notar problemas relacionados às informações que eram fornecidas por autores que submetiam seus trabalhos.

Por fim, um experimento foi conduzido a fim de avaliar a proposta e compara-la em termos de eficácia com outras ferramentas através das métricas de precisão, revocação e medida-f. Essas métricas foram avaliadas em dois cenários de alinhamento, nos quais a proposta obteve primeiro e segundo lugar. Apesar de não ter tido os melhores valores em ambas as avaliações, a proposta apresentada se destaca pela ausência de implementações específicas para o *dataset*, gerando menos impacto quando é necessária uma mudança de contexto.

Principais contribuições

As principais contribuições deste trabalho são apresentadas a seguir:

- Desenvolvimento de processo alinhamento de dados conectados, permitindo que *datasets* reais possam ser conectados semi-automaticamente;
- Abordagem capaz de calcular a similaridade de recursos levando em consideração o modelo ontológico;
- Viabilização da execução do alinhamento diretamente dentro do armazenamento de triplas;
- Criação de experimento e estudo de caso para avaliar a eficácia das soluções de alinhamento no estado da arte.

Em linhas gerais, tem-se que esta proposta pode ser aplicada a diferentes domínios, permitindo que seja possível extrair contribuições em cada um deles:

- **Contribuições para Informática na Educação (estudo de caso):** foi possível gerar um panorama da comunidade, permitindo que através do cruzamento de dados, a comunidade

tenha uma visão mais geral e através das respostas fornecidas será possível adotar decisões estratégicas.

- **Dados Conectados:** A partir da utilização do processo, é possível construir uma abordagem livre de contexto para a correspondência de instâncias. Essa abordagem pode ser aplicada em processos de publicação de dados conectados, permitindo que os dados sejam enriquecidos durante a publicação. Além disso, o consumo de dados conectados pode explorar o processo, visto que os dados podem ser alinhados com possíveis *datasets* locais.
- **Governo:** A abordagem permite que os dados sejam enriquecidos através da correspondência entre instâncias, contribuindo com iniciativas de interoperabilidade de dados.
- **Cidadão** Através das correspondências geradas, a sociedade poderia identificar e monitorar com maior facilidade entidades públicas de seu interesse.

Limitações e trabalhos futuros

Algumas questões que não foram foco deste trabalho, mas que devem ser consideradas em trabalhos futuros, são estudos relacionados à qualidade e o tamanho (quantidade de triplas) do *dataset* impacta na qualidade dos alinhamentos.

Outra questão que não está dentro do escopo deste trabalho, mas que também deve ser levado em consideração, é o alinhamento entre ontologias, pois este trabalho foi desenvolvido com objetivo de alinhar dados. Portanto, faz-se necessário estudar mecanismos que unifiquem soluções para alinhar dados e alinhar ontologias.

Com trabalhos futuros, pretende-se realizar mais experimentos para analisar a eficácia da ferramenta com *datasets* com diversas características (domínio, quantidade de triplas, qualidade etc.). Além disso, pretende-se disponibilizar a solução como uma infraestrutura de alinhamento disponível na Web, dando acesso aos alinhamentos realizados através da solução. Por fim, outras pesquisas serão aplicadas com os seguintes objetivos:

- **Automatizar o processo de alinhamento:** Um possível caminho para isso seria a escolha automática dos conceitos relacionados;
- **Otimizar a performance:** Como os conceitos relacionados podem ser alinhados paralelamente, possíveis abordagens seriam paralelismos e distribuição.
- **Melhorar a qualidade no cálculo de similaridade entre recursos:** Uma abordagem possível seria a composição de funções de similaridade e a identificação de características que apresentem maior expressividade na identificação de similaridade.

REFERÊNCIAS

ARAUJO, S. et al. Serimi: resource description similarity, rdf instance matching and interlinking. In: CEUR-WS. ORG. *Proceedings of the 6th International Conference on Ontology Matching-Volume 814*. [S.l.], 2011. p. 246–247. 15

BERNERS-LEE, T. Linked data-design issues. 2006. 10, 19, 20

BERNERS-LEE, T. et al. The semantic web. *Scientific american*, New York, NY, USA:, v. 284, n. 5, p. 28–37, 2001. 10

BIZER, C. et al. How to publish linked data on the web. 2007. 10, 18

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, p. 205–227, 2009. 10

BORST, W. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. Tese (Doutorado), 1997. 19

BRINKER, S. Best buy jump starts data web marketing. *Chief marketing technologist*, v. 11, 2009. 10

CASTANO, S. et al. Ontology and instance matching. In: *Knowledge-driven multimedia information extraction and ontology evolution*. [S.l.]: Springer, 2011. p. 167–195. 11, 12, 36, 39

CHEATHAM, M. et al. Results of the ontology alignment evaluation initiative 2015. In: NO COMMERCIAL EDITOR. *10th ISWC workshop on ontology matching (OM)*. [S.l.], 2015. p. 60–115. 13, 24

DEVED, V. et al. *Semantic web and education*. [S.l.]: Springer Science & Business Media, 2006. v. 12. 19

FARIA, D. et al. Oaei 2016 results of aml. 2016. 16, 24

FERRARA, A. et al. Towards a benchmark for instance matching. In: CEUR-WS. ORG. *Proceedings of the 3rd International Conference on Ontology Matching-Volume 431*. [S.l.], 2008. p. 37–48. 13, 23, 24, 36

FISHER, R. A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, JSTOR, v. 222, p. 309–368, 1922. 46

GOUTTE, C.; GAUSSIER, E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2005. p. 345–359. 54

GRUBER, T. R. et al. A translation approach to portable ontology specifications. *Knowledge acquisition*, Citeseer, v. 5, n. 2, p. 199–220, 1993. 19

GU, L. et al. Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report*, Citeseer, v. 3, p. 83, 2003. 11

HOLANDA, O. et al. Joint: Java ontology integrated toolkit. *Expert Systems with Applications*, Elsevier, v. 40, n. 16, p. 6469–6477, 2013. 38

HOMOCEANU, S.; KALO, J.-C.; BALKE, W.-T. Putting instance matching to the test: Is instance matching ready for reliable data linking? In: SPRINGER. *International Symposium on Methodologies for Intelligent Systems*. [S.l.], 2014. p. 274–284. 12, 13, 23, 39

HYLAND, B.; WOOD, D. The joy of data-a cookbook for publishing linked government data on the web. In: *Linking government data*. [S.l.]: Springer, 2011. p. 3–26. 10, 18

ISOTANI, S.; BITTENCOURT, I. I. *Dados Abertos Conectados*. [s.n.], 2015. 175 p. ISSN 24470821. ISBN 978-85-7522-449-6. Disponível em: <http://ceweb.br/livros/dados-abertos-conectados/>. 10, 19, 20

JIMENEZ, S. et al. Generalized mongue-elkan method for approximate text string comparison. In: SPRINGER. *International Conference on Intelligent Text Processing and Computational Linguistics*. [S.l.], 2009. p. 559–570. 22

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions and reversals. In: *Soviet physics doklady*. [S.l.: s.n.], 1966. v. 10, p. 707. 20, 37

LI, J. et al. Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 21, n. 8, p. 1218–1232, 2009. 25

MENESTRINA, D.; BENJELLOUN, O.; GARCIA-MOLINA, H. Generic entity resolution with data confidences. Stanford, 2005. 11

MONGE, A. et al. The field matching problem: Algorithms and applications. In: *KDD*. [S.l.: s.n.], 1996. p. 267–270. 21, 22

MONTGOMERY, D. C. *Design and analysis of experiments*. 8th. ed. [S.l.]: John Wiley & Sons, 2012. 42

NGUYEN, K.; ICHISE, R.; LE, B. Interlinking linked data sources using a domain-independent system. In: SPRINGER. *Joint International Semantic Technology Conference*. [S.l.], 2012. p. 113–128. 13

ROYCE, W. W. Managing the development of large software systems. In: LOS ANGELES. *proceedings of IEEE WESCON*. [S.l.], 1970. v. 26, n. 8, p. 328–338. 33

SARAWAGI, S.; BHAMIDIPATY, A. Interactive deduplication using active learning. In: ACM. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2002. p. 269–278. 11

SHADBOLT, N. et al. Linked open government data: Lessons from data. gov. uk. *IEEE Intelligent Systems*, v. 27, n. 3, p. 16–24, 2012. 10

SIMISTER, S.; BRICKLEY, D. Simple application-specific constraints for rdf models. In: *RDF Validation Workshop. Practical Assurances for Quality RDF Data, Cambridge, Ma, Boston*. [S.l.: s.n.], 2013. 10

SINGHAL, A. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, v. 24, n. 4, p. 35–43, 2001. 37

SOMMERVILLE, I. *Software Engineering*. Pearson, 2011. (International Computer Science Series). ISBN 9780137053469. Disponível em: <https://books.google.ca/books?id=l0egcQAACAAJ>. 33

STUDER et al. Knowledge engineering: principles and methods. *Data & knowledge engineering*, Elsevier, v. 25, n. 1, p. 161–197, 1998. 19

VILLAZÓN-TERRAZAS, B. et al. Methodological guidelines for publishing government linked data. In: *Linking government data*. [S.l.]: Springer, 2011. p. 27–49. 10, 18

WANG, P.; XU, B. Lily: Ontology alignment results for oaei 2008. In: CEUR-WS. ORG. *Proceedings of the 3rd International Conference on Ontology Matching-Volume 431*. [S.l.], 2008. p. 167–175. 37

WINKLER, W. E. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. ERIC, 1990. 37

WOOD, D. et al. *Linked Data*. [S.l.]: Manning Publications Co., 2014. 10

ZHANG, Y. et al. Rimom results for oaei 2016. p. 7, 2016. 16, 25, 53

ÁVILA, T. J. T. *Uma proposta de modelo de processo para publicação de dados abertos conectados governamentais*. 223 p. Dissertação (Master Thesis) — Universidade Federal de Alagoas, 2015. 10, 18