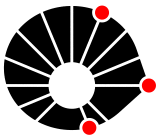
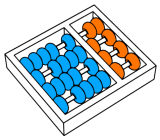




Refinamento de Mapeamentos dirigido pela Evolução de Ontologias

Victor Eiti Yamamoto *Julio Cesar dos Reis*

Technical Report - IC-19-07 - Relatório Técnico
September - 2019 - Setembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Refinamento de Mapeamentos dirigido pela Evolução de Ontologias

Victor Eiti Yamamoto
Julio Cesar dos Reis
Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP, Brasil.

2019

Resumo

Ontologias e seus mapeamentos associados exercem papel central em diversas atividades relacionadas com interoperabilidade semântica em sistemas computacionais. Contudo, as evoluções constantes das ontologias exigem atualizações dos alinhamentos existentes. Apesar das técnicas de manutenção de mapeamentos terem lidado com revisões e remoções, a adição de conceitos demandam mais estudos. Esta pesquisa propõe técnicas para refinar um conjunto de mapeamentos estabelecidos com base na evolução de ontologias. Investigamos meios de sugerir correspondências em uma nova versão da ontologia sem aplicar operações de correspondência com todo o conjunto de entidades da ontologia. Os resultados obtidos exploram a vizinhança dos conceitos no processo de alinhamento para atualizar o conjunto de mapeamentos. Nossa avaliação experimental com diversas versões de alinhamentos entre ontologias biomédicas evidenciam a efetividade em considerar o contexto de novos conceitos.

1 Introdução

Nas últimas décadas, a área biomédica explorou ontologias e suas capacidades para vários propósitos como coleta, administração e compartilhamento de informações. Contudo, o tamanho desse domínio exige o uso de várias ontologias que são ligadas por meio de mapeamentos. Os mapeamentos são a materialização das relações semânticas entre elementos das ontologias inter-relacionadas [14].

Criar mapeamentos entre ontologias é uma atividade complexa, especialmente por causa do crescimento das ontologias biomédicas. Várias técnicas automáticas de alinhamento foram propostas [9]. Entretanto, um significativo esforço manual para

validação é exigido se um certo nível de qualidade é requerida. Isso impede aplicações de confiarem no mapeamento para obter total proveito deles.

Ontologias nas ciências da vida evolui para se manter atualizado com o domínio do conhecimento. Mudanças na ontologia pode afetar mapeamentos existentes ou pode auxiliar no processo de refinamento dos mesmos. Nesse contexto, para evitar o custoso processo de realinhamento de ontologias, é necessário utilizar uma estratégia para adaptar os mapeamentos com o objetivo de mantê-los válidos semanticamente [11]. Realizar manutenção manual dos mapeamentos é possível apenas quando as modificações são aplicadas em uma quantidade restrita de mapeamentos. Do contrário, é necessário utilizar métodos automáticos para ontologias grandes e dinâmicas. Ontologias biomédicas costumam conter uma quantidade alta de conceitos interconectados por mapeamentos.

Para lidar com o problema de reconciliação de mapeamentos de forma semi-automática algumas questões devem ser resolvidas. Primeiramente, é difícil avaliar o real impacto da evolução de ontologias nos mapeamentos. Nesta situação, o problema está em identificar e classificar os diferentes casos. Segundo, diferentes tipos de alteração podem ser aplicadas nas ontologias, mas ainda é necessário investigar quais tipos de alteração devem ser consideradas para a reconciliação de mapeamento [3].

Trabalhos anteriores apresentaram uma estratégia de adaptação de mapeamentos para dois de três categorias de evolução: *remoção* de conhecimento e *revisão* de conhecimento [11]. Por exemplo, quando um conceito é removido, heurísticas foram desenvolvidas para aplicar as adaptações necessárias no mapeamento. A *adição* de conhecimento (terceira categoria) é o tipo de alteração mais frequente nas evoluções de ontologias. Novos conceitos ou atributos dos conceitos são adicionados para adequar-se com os novos domínios de conhecimento. Esses novos conhecimentos devem ser alinhadas com as ontologias inter-relacionadas.

Neste relatório descrevemos um método de refinamento de alinhamento de ontologias para atualizar um conjunto de mapeamento considerando as mudanças ocorridas nas ontologias (baseado nos novos conceitos adicionados na evolução de ontologias). Estudamos o uso de informações relacionados com a vizinhança dos conceitos para melhorar a qualidade dos mapeamentos obtidos. Para essa finalidade, investigamos técnicas para reutilizar os mapeamentos já estabelecidos e explorar a função das vizinhanças dos conceitos para obter novos mapeamentos. Nossa proposta permite sugerir novas correspondências sem aplicar a operação de correspondência com todo o conjunto de entidades da ontologia.

Nossa avaliação experimental explora ontologias biomédicas reais e mapeamentos estabelecidos entre elas. O conjunto de mapeamentos criados com base nas sugestões geradas automaticamente foram comparados com o conjunto de novas correspondências observadas nas versões atualizadas dos mapeamentos e avaliamos a sua qualidade por meio de métricas de avaliação padrão. Demonstramos que a corres-

pondência considerando a vizinhança dos conceitos é competitivo com as operações de correspondência com todo o conjunto da ontologia alvo.

O restante deste relatório está organizado da seguinte forma: na seção 2 apresentamos uma revisão sintética da literatura; a seção 3 formaliza os conceitos fundamentais no contexto da pesquisa; a seção 4 descreve a técnica de refinamento de alinhamento entre ontologias com base em novos conceitos; a seção 5 apresenta os experimentos realizados e seus resultados; a seção 6 elabora uma discussão sobre os resultados enquanto a seção 7 efetua as conclusões finais.

2 Revisão da Literatura

Estudos anteriores investigou métodos semi-automáticos para adaptar mapeamentos de ontologias quando ao menos uma das ontologias mapeadas evolui [3]. Dos reis *et al.* conceituou *DyKOSMap* framework [11] para ajudar na adaptação de mapeamento semântico destacando diferentes aspectos como: a função de diferentes tipos de alteração de ontologias, a importância em considerar as informações dos conceitos que os mapeamentos estabelecidos estão relacionados, além da relevância dos diferentes tipos de relações semânticas dos mapeamentos.

Algumas técnicas utilizaram recursos externos almejando melhorar e aumentar o número e a precisão dos mapeamentos estabelecidos. Stoutenburg [15] argumentou que o uso de ontologias superiores (uma ontologia consistido de termos genéricos que são comuns pelo todo o domínio) e recursos linguísticos pode melhorar o processo de alinhamento.

A ferramenta de correspondência *TaxoMap* [5] explorou técnicas de mapeamento baseado em padrões. O mapeamento é gerado com relações iniciais propostas (correspondências encontradas são relações equivalentes, relações de subsunção e seu inverso ou relações de proximidade). Um especialista no domínio valida manualmente os mapeamentos gerados e corrige os problemas, agrupando os problemas identificados quando eles correspondem a casos similares. A ferramenta gera padrões baseado nos grupos de casos similares que podem ser aplicado para outros mapeamentos no mesmo domínio.

Outras técnicas combinaram algoritmos baseados na semântica e no léxico, a maioria utilizando recursos disponíveis no *Unified Medical Language System* (UMLS) [1] para gerar mapeamentos. O uso de UMLS como um recurso externo pode ser interessante em vários aspectos: (1) favorece o aumento no número de mapeamentos, (2) providencia diferentes termos sinônimos para um dado conceito e (3) define relações entre conceitos em uma rede semântica. Zhand e Bodenreider [16] explorou UMLS para melhorar o alinhamento entre ontologias anatômicas. Eles mostraram que o

[1] UMLS é uma coleção de vocabulários de saúde e biomédico e normas URL: www.nlm.nih.gov/research/umls/

domínio de conhecimento é um fator chave para a identificação de mapeamentos adicionais comparado com métodos utilizando esquemas genéricos de mapeamento.

Sekhavat e Parsons [13] exploraram modelos conceituais (*e.g.*, Relações de entidades, diagramas de classes ou domínio das ontologias) como um conhecimento prévio para enriquecer os mapeamentos em banco de dados e resolver mapeamentos ambíguos. Nesta abordagem utilizou modelos de conceitos como recursos externos para capturar semântica de elementos do esquema, por exemplo, um par de conceitos a_1 e a_2 onde a_1 é uma subclasse e a_2 é uma superclasse no modelo conceitual. Esta informação foi utilizada para enriquecer o esquema antes do mapeamento, marcando as chaves estrangeiras correspondentes a a_1 e a_2 como generalizações. Como consequência, os relacionamentos identificados no mapeamento de esquema é uma generalização (*é–um*) no lugar de equivalência.

Pruski *et al.* [10] propôs explorar fontes externas específicas do domínio para caracterizar a evolução dos conceitos em ontologias dinâmicas. A técnica analisou a evolução dos valores nos atributos dos conceitos. Esta abordagem utilizou propriedades das ontologias e mapeamentos entre ontologias de repositórios online para deduzir o relacionamento entre os conceitos e suas sucessivas versões.

Noy *et al.* [8] e Seddiqui *et al.* [12] explorou conceitos âncoras para obter mapeamentos. Eles utilizaram um conjunto de pares de conceitos alinhados para obter outros mapeamentos baseados nesses pares. A abordagem calcula novos alinhamentos para todos os conceitos das ontologias envolvidas, portanto eles não são utilizados para evolução de ontologias.

Nesta pesquisa, exploramos as operações de alteração de ontologias para alavancar refinamentos, em particular, conceitos adicionados. Contribuímos com uma metodologia para considerar novos conceitos adicionados e investigamos o contexto dos conceitos candidatos de mapeamentos existentes para refinar ao longo do tempo. Avaliamos o algoritmo proposto para medir a efetividade dos nossos mapeamentos refinados com ontologias biomédicas reais.

3 Fundamentos e Formalizações

Ontologias. Uma ontologia $\mathcal{O}$ especifica a conceitualização de um domínio em termos de conceitos, atributos e relacionamentos [4]. Formalmente, uma ontologia $\mathcal{O} = (\mathcal{C}_\mathcal{O}, \mathcal{R}_\mathcal{O}, \mathcal{A}_\mathcal{O})$ consiste de um conjunto de conceitos $\mathcal{C}_\mathcal{O}$ inter-relacionados por um relacionamento direcionado $\mathcal{R}_\mathcal{O}$. Cada conceito $c \in \mathcal{C}_\mathcal{O}$ tem um identificador único e é associado com um conjunto de atributos $\mathcal{A}_\mathcal{O}(c) = \{a_1, a_2, ..., a_p\}$. Cada relacionamento $r(c_1, c_2) \in \mathcal{R}_\mathcal{O}$ é tipicamente uma tripla (c_1, c_2, t) onde t é uma relação (*e.g.*, "é_um", "parte_de", "ajudado_por", *etc.*) inter-relacionando c_1 e c_2.

Contexto de um conceito. O contexto de um conceito $c_i \in \mathcal{C}_\mathcal{O}$ é definidido como um conjunto de *super conceitos*, *sub conceitos* e *conceitos irmãos* de c_i.

$$CT(c_i, \lambda) = sup(c_i, \lambda) \cup sub(c_i, \lambda) \cup sib(c_i, \lambda) \tag{1}$$

onde

$$\begin{array}{ll} sup(c_i, \lambda) & = \{c_j | c_j \in \mathcal{C}_\mathcal{O}, r(c_i, c_j) = \text{“} \sqsubset \text{”} \wedge distância(c_i, c_j) \leq \lambda \wedge c_i \neq c_j\} \\ sub(c_i, \lambda) & = \{c_j | c_j \in \mathcal{C}_\mathcal{O}, r(c_j, c_i) = \text{“} \sqsubset \text{”} \wedge distância(c_i, c_j) \leq \lambda \wedge c_i \neq c_j\} \\ sib(c_i, \lambda) & = \{c_j | c_j \in \mathcal{C}_\mathcal{O}, ((sup(c_j) \cap sup(c_i)) \vee (sub(c_j) \cap sub(c_i))) \\ & \wedge distância(c_i, c_j) \leq \lambda \wedge c_i \neq c_j\} \end{array} \tag{2}$$

onde λ é o nível do contexto. O nível do contexto define o valor máximo para a distância entre dois conceitos (em termos da menor distância de relacionamento na hierarquia de conceitos) e o símbolo "$\sqsubset$" indica que "c_i é um sub conceito de c_j". Esta definição de $CT(c_i, \lambda)$ define os conceitos relevantes a serem investigados no refinamento de mapeamento.

Similaridade entre conceitos. Dado dois conceitos particulares c_i e c_j, a similaridade entre eles pode ser definido como a similaridade máxima entre cada dupla de atributos do c_i e c_j. Formalmente:

$$sim(c_i, c_j) = \arg\max \ sim(a_{ix}, a_{jy}) \tag{3}$$

onde $sim(a_{ix}, a_{jy})$ é a similaridade entre dois atributos a_{ix} e a_{jy} denotando os conceitos c_i and c_j, respectivamente.

Mapeamento. Dado dois conceitos c_s e c_t de duas ontologias diferentes, um mapeamento m_{st} pode ser definido como:

$$m_{st} = (c_s, c_t, semType, conf) \tag{4}$$

onde $semType$ é a relação semântica conectando c_s e c_t. Neste artigo, diferenciamos *relação* e *relacionamento*, onde o primeiro se refere a mapeamentos e o segundo para as ontologias. Os seguintes tipos de relações semânticas são consideradas: *não mapeável* [$\perp$], *equivalente* [$\equiv$], *específico para genérico* [$\leq$], *genérico para específico* [$\geq$] e *sobreposição* [$\approx$]. Por exemplo, conceitos podem ser equivalentes (*e.g.*, "cabeça"$\equiv$"cabeça"), um conceito pode ser mais ou menos genérico do que outros (*e.g.*, "polegar"$\leq$"dedo") ou os conceitos podem ser semanticamente relacionados ($\approx$). O $conf$ é a similaridade entre c_s e c_t indicando a confiança de sua relação [2]. Nós definimos $\mathcal{M}^j_{ST}$ como um conjunto de mapeamentos m_{st} entre ontologias $\mathcal{O}_S$ e $\mathcal{O}_T$ em um dado tempo j. Nós assumimos $j \in N$ a versão da ontologia $\mathcal{O}^j_S$. Ontologia $\mathcal{O}^0_S$ é a versão 0 e $\mathcal{O}^1_S$ é a versão 1 da mesma ontologia.

Operações de alteração de ontologias (OCO). Uma operação de alteração de ontologias (OCO) é definido para representar as mudanças em um atributo, em um conjunto de um ou mais conceitos ou numa relação entre conceitos. OCOs são classificadas em duas categorias principais: Alterações *atômicas* e *complexas*. Cada

operação da primeira categoria não pode ser divida em operações menores, enquanto o segundo é composta de mais de uma operação atômica. Neste artigo, focamos na operação de adição de conceitos que é uma operação atômica.

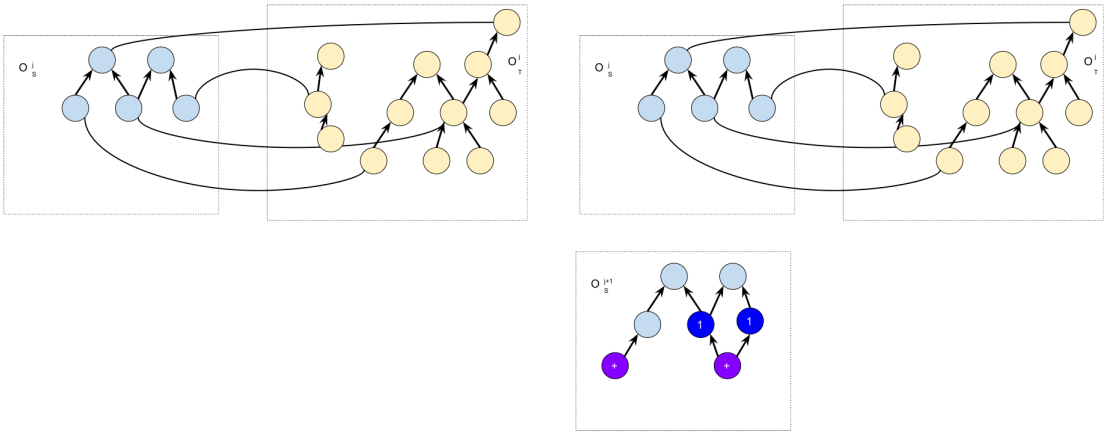
Caracterização do problema. Considerando duas versões da mesma ontologia fonte $\mathcal{O}_S^j$ no tempo j e $\mathcal{O}_S^{j+1}$ no tempo $j+1$, uma ontologia alvo $\mathcal{O}_T^j$ e um conjunto de mapeamento inicial $\mathcal{M}_{ST}^j$ entre $\mathcal{O}_S^j$ e $\mathcal{O}_T^j$ no tempo j. Suponha que a frequência de novas versões de $\mathcal{O}_S$ e $\mathcal{O}_T$ é diferente e no tempo $j+1$ apenas $\mathcal{O}_S$ evolui. Como a evolução deve impactar no mapeamento $\mathcal{M}_{ST}^j$, é necessário refinar $\mathcal{M}_{ST}^j$ para garantir a qualidade e a completude de $\mathcal{M}_{ST}^{j+1}$. A qualidade é relacionada com a consistência do mapeamento e pode ser medido usando a precisão. Por exemplo, mapeamentos não podem ser estabelecidas entre conceitos removidos. A completude se refere a cobertura dos conceitos alinhados em $\mathcal{M}_{ST}^{j+1}$. Neste trabalho, estudamos como $\mathcal{M}_{ST}^j$ pode ser refinado baseado nas alterações da ontologia relacionado com *adição de conhecimento*. Nós abordamos as seguintes questões:

- Como explorar mapeamentos existentes para refinar os mapeamentos baseado nos novos conceitos adicionados?
- É possível refinar mapeamentos para alinhar os novos conceitos sem realizar a operação de correspondência em toda ontologia alvo?
- Qual é o impacto na efetividade ao utilizar o contexto dos conceitos $CT(c_i, \lambda)$ nas ontologias fonte e alvo no refinamento de mapeamento?

Consideramos que $\mathcal{O}_T$ não evoluiu (dessa forma $\mathcal{O}_T^j$ e $\mathcal{O}_T^{j+1}$ estão na mesma versão da ontologia $\mathcal{O}_T$).$\mathcal{O}_S^j$ e $\mathcal{O}_S^{j+1}$ são duas versões distintas da mesma ontologia $\mathcal{O}_S$. No tempo $j+1$, novos conceitos adicionados aparecem em $\mathcal{O}_S^{j+1}$ e objetivamos refinar o conjunto de mapeamento original M_{ST}^j para obter um conjunto de mapeamentos validos e atualizados em M_{ST}^{j+1}.

4 Refinamento de alinhamento entre ontologias com base em novos conceitos

Nosso objetivo é propor correspondências adequadas para cada novo conceito adicionado no tempo $j+1$. No primeiro passo, nossa abordagem identifica todos os novos conceitos adicionados utilizando a ferramenta *Conto-Diff* [6]. Esta ferramenta permite identificar alterações atômicas e complexas na ontologia. Depois, extraímos a informação contextual, *i.e.* super, sub e conceitos irmãos para esses novos conceitos adicionados (*cf.* Formula 1). Então examinamos o mapeamento entre o conceito fonte no contexto dos novos conceitos adicionados e seu correspondente conceito alvo. A



(a) Situação inicial (b) Encontrando o contexto na ontologia fonte

Figura 1: Representação das situações antes de aplicar o algoritmo de alinhamento

ideia por trás da técnica orientado a contexto é que o mapeamento candidato é estabelecida entre um novo conceito adicionado e um conceito alvo de um mapeamento existente no tempo t.

A Figura 1 (a) ilustra uma situação onde existem duas ontologias que tem um alinhamento no tempo j. Cada círculo representa um conceito de uma ontologia. Os círculos azul claro representam conceitos da ontologia fonte. Os círculos amarelos representam os conceitos da ontologia alvo. Linhas contínuas representam mapeamento entre conceitos da ontologia fonte e ontologia alvo.

A Figura 1 (b) ilustra uma situação em que a ontologia fonte evoluiu e mudou para o tempo $j + 1$. O algoritmo encontra os novos conceitos adicionados e explora o contexto de cada novo conceito adicionado. Neste caso, exploramos o contexto do conceito direito utilizando nível fonte 1. Os círculos roxos representam novos conceitos adicionados. Os círculos azuis escuros representam conceitos de um contexto com um certo nível fonte; o número dentro do círculo representam o nível fonte necessário para acessar esse conceito.

Após encontrar alguns conceitos dentro do contexto dos novos conceitos adicionados que tem um alinhamento no tempo anterior, os conceitos da ontologia alvo que tem um alinhamento no tempo anterior são adicionados como conceitos candidatos. O contexto de cada conceito candidato é explorado e adicionado como conceito candidato. A Figura 2 ilustra a situação usando nível alvo 1. Círculos vermelhos representam conceitos candidatos para o novo conceito na ontologia fonte; o número dentro do círculo representa o nível alvo que representa o nível alvo necessario para acessar o conceito. As linhas pontilhadas representam os possíveis alinhamentos entre o novo conceito (em $\mathcal{O}_S^{j+1}$) e conceitos candidatos (em $\mathcal{O}_T^j$).

O Algoritmo 1 computa o *diff* entre duas versões da ontologia fonte (linha 1). Para

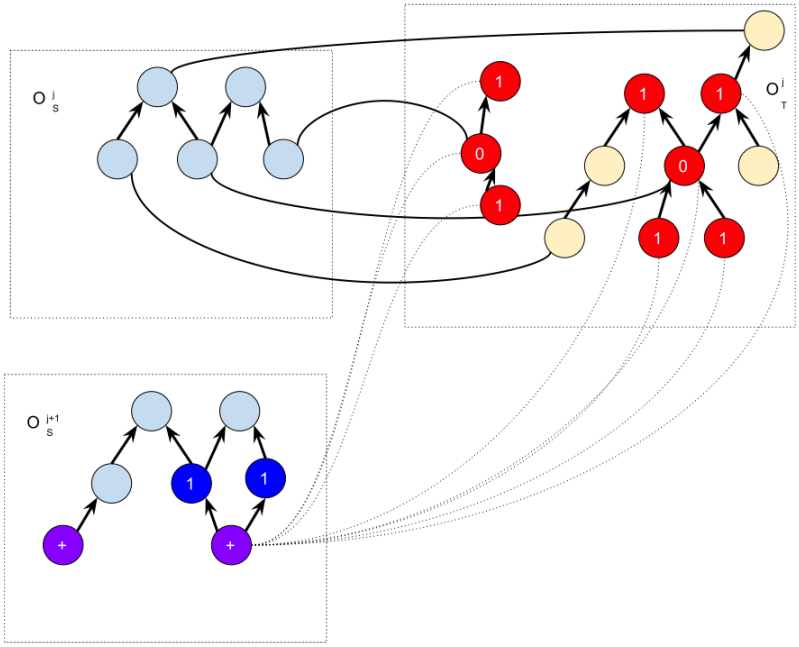


Figura 2: Calculando similaridade com conceitos candidatos

cada novo conceito adicionado c_i^1, o algoritmo considera um conceito candidato c_t^0 na ontologia alvo ao explorar os mapeamentos já existentes relacionados com $CT(c_i^1, \gamma)$ (linhas 4-8). Note que recuperamos a versão anterior à evolução (c_k^0) do conceito c_k^1 encontrado no contexto de c_i^1.

Para cada c_t^0, o algoritmo obtêm um conjunto de conceitos da $CT(c_t^0, \lambda)$ (linha 11). Determinamos um novo mapeamento refinado ao calcular a similaridade entre um novo conceito c_i^1 da $\mathcal{O}_S^{j+1}$ e um candidato $c_n \in \mathcal{C}_t$. Se o maior valor de similaridade (entre os atributos dos conceitos) é igual ou maio que o limiar τ, o algoritmo estabelece um mapeamento entre os novos conceitos adicionados e o conceito candidato do alvo. Algoritmo 1 pesquisa pelos candidatos c_t que retorna o valor de maior similaridade.

Para comparar com os resultados obtidos pela nossa abordagem (*cf.* Seção 5), propusemos outro algoritmo que ignora o contexto dos novos conceitos para calcular a similaridade. Isso significa que o algorítimo computa a similaridade entre cada novo conceito com todos os conceitos na ontologia alvo. Mais especificamente, o algoritmo computa o *diff* entre duas versões da ontologia alvo. Para cada novo conceito adicionado, ele calcula a similaridade entre todos os conceitos com a ontologia alvo. Se existe qualquer similaridade maior que o limiar, o algoritmo cria um novo mapeamento entre os novos conceitos adicionados e um conceito na ontologia alvo com o maior valor de similaridade.

No nosso algoritmo, os atributos dos conceitos alvo são comparados com todos os atributos do conceito alvo para obter a similaridade entre os conceitos. O valor de

Algorithm 1: Abordagem contextual de refinamento de mapeamento

Require: $\mathcal{O}_S^j, \mathcal{O}_S^{j+1}, \mathcal{O}_T^j, \mathcal{O}_T^{j+1}, \mathcal{M}_{ST}^j, \lambda, \gamma, \tau \in \mathbb{R}$
Ensure: $M_A = \{m_1, m_2, ..., m_N\}$
1: $\mathcal{C}_{add} \leftarrow diff_{add}(\mathcal{O}_S^j, \mathcal{O}_S^{j+1})$ {novos conceitos adicionados}
2: $\mathcal{C}_t \leftarrow \emptyset$ {Initializa conceitos alvos do mapeamento candidato}
3: **for all** $c_i^1 \in \mathcal{C}_{add}$ **do**
4: **for all** $c_k^1 \in CT(c_i^1, \gamma)$ **do**
5: **if** $\exists c_t^0 \in \mathcal{C}_{\mathcal{O}_T^0}, \exists m(c_k^0, c_t^0) \in \mathcal{M}_{ST}^0$ **then**
6: $\mathcal{C}_t \leftarrow \mathcal{C}_t \bigcup \{c_t^0\}$
7: **end if**
8: **end for**
9: $m_{it} \leftarrow \emptyset$
10: **for all** $c_t \in \mathcal{C}_t$ **do**
11: **for all** $c_n \in CT(c_t, \lambda)$ **do**
12: $m_{cand} \leftarrow$ argmax $sim(c_i^1, c_n)$ {Criando um mapeamento entre conceitos c_i^1 e c_n}
13: **if** $\max(sim(c_i^1, c_n)) \geq \tau$ **then**
14: $m_{it} \leftarrow m_{cand}$
15: $\tau \leftarrow \max(sim(c_i^1, c_n))$
16: **end if**
17: **end for**
18: **end for**
19: $\mathcal{M}_A \leftarrow \mathcal{M}_A \bigcup \{m_{it}\}$
20: **end for**
21: **return** $\mathcal{M}_A$

similaridade entre dois conceitos é o maior valor de similaridade entre os atributos. O método utilizado para calcular a similaridade afeta a precisão e a cobertura. Neste trabalho, exploramos o Bi-grama Dice para calcular a similaridade. Bi-grama é uma sequência de duas letras adjacente de uma palavra. O coeficiente de Dice é definido como duas vezes a quantidade de elementos comuns dividido pela soma de cada elemento. A formula 5 apresenta a aplicação do bi-grama Dice das strings X e Y. A vantagem do n-grama está no fato de ser sensível ao contexto, mas não tem uma boa resolução quando o tamanho do grama é aumentado [7].

$$Similaridade = \frac{2 \times (Bi - grama(X) \cap Bi - grama(Y)}{Bi - grama(X) + Bi - grama(Y)} \tag{5}$$

5 Experimentos

O objetivo na avaliação é averiguar a qualidade do conjunto de mapeamento refinado como resultado da nossa abordagem. Dados utilizados nessa avaliação foram obtidos de cinco ontologias biomédicas: SNOMED-CT (SCT), MeSH, ICD-9-CM, ICD10-CM e NCI Thesaurus. SNOMED-CT (*Systematized Nomenclature of Medicine—Clinical Terms*) é uma ontologia no qual o objetivo é criar uma taxonomia de termos referente ao ambiente médico e uma framework de regras garantindo que cada termo seja utilizado com exatamente um significado [1]. *MeSH Thesaurus* é um vocabulário controlado produzido pela N*ational Library of Medicine* e utilizado para indexar, catalogar e pesquisar por informações e documentos relacionados a biomedicina e saúde urlhttps://www.nlm.nih.gov. ICD-9-CM e ICD-10-CM são formalizações em OWL-DL da Classificação Internacional de Doenças e Problemas Relacionados com a Saúde publicado pela Organização Mundial da Saúde[2]. *NCI Thesaurus* [3] contém as terminologias utilizadas na infraestrutura semântica e sistemas de informação da *National Cancer Institute*. A Tabela 2 apresenta estatísticas sobre as ontologias fontes e alvos para cada versão considerada.

Ontologias	**Versão**	**#Conceitos**	**#Atributos**	**#Relação é-um**	**#Novos conceitos**
ICD9	2009	12.734	34.065	11.619	325
	2011	13.059	34.963	11.962	
ICD10	2011	43.351	87.354	40.330	0
SCT	2010	386.965	1.531.288	523.958	8.381
	2012	395.346	1.570.504	539.245	
NCI	2009	77.448	282.434	86.822	17.284
	2012	94.732	365.515	105.406	
MeSH	2012	50.367	259.565	59.191	604
	2013	50.971	264.783	59.844	

Tabela 2: Estatística das ontologias

Os mapeamentos obtidos pelo Algoritmo 1 são comparados com os mapeamentos oficiais. Os mapeamentos entre SNOMEDCT e ICD9CM é oferecida pelo International Health Terminology Standards Development Organisation (IHTSDO)[4]. Os mapeamentos entre MeSH e ICD-10-CM são oferecidos pelo *Catalogue et Indexation des Sites Médicaux de langue Française* (CISMeF)[5]. A Tabela 3 apresenta a quantidade de cada conjunto de mapeamento entre as ontologias utilizadas neste experimento.

Para analisar os resultados obtidos experimentalmente, é necessário comparar os mapeamentos obtidos pelo algoritmo com os mapeamentos criados apenas para os novos conceitos adicionados na nova versão das ontologias consideradas. A Tabela 4 apresenta a quantidade de mapeamentos considerados na métrica.

[2]http://www.who.int/classifications/icd/en/

[3]https://ncit.nci.nih.gov/ncitbrowser/

[4]ttps://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html

[5]http://www.chu- rouen.fr/cismef

SCT-ICD9	**#Mapeamentos**	**SCT-NCI**	**#Mapeamentos**	**MeSH-ICD10CM**	**#Mapeamentos**
2010-2009	84,519	2009-2009	19,971	2012-2011	4,631
2012-2011	86,638	2012-2012	22,732	2013-2011	5,378

Tabela 3: Estatística dos mapeamentos estudados

Mapeamentos	**#Mapeamentos oficiais criados para os novos conceitos adicionados**
SNOMEDCT-ICD9CM	1,583
SNOMEDCT-NCI	158
MeSH-ICD10CM	21

Tabela 4: Número de novos mapeamentos criados e associados com novos conceitos adicionados na nova versão da ontlogia

O experimento foi realizado para três conjunto de dados (SCT-NCI, SCT-ICD9 e MeSH-ICD10) considerando SCT e MeSH como ontologias fontes. Para avaliar as configurações, consideramos três níveis fontes, três valores de limiar (0.5, 0.75 e 0.9) e quatro nível alvo. Para cada conjunto de dados, fixamos o nível fonte e o limiar para verificar o resultado para cada nível fonte. Depois de examinar todos os níveis alvo, alteramos o limiar e repetimos para cada nível fonte. Após examinar para cada valor de limiar, alteramos o nível fonte e repetimos a operação com todos os valores de limiar e nível alvo.

Utilizamos três métricas para avaliar os resultados: Precisão, cobertura e F-Measure. Essas métricas foram utilizadas comparando os resultados obtidos pela nossa abordagem e os resultados esperados pelo mapeamento oficial.

Precisão é definido como a relação entre os mapeamentos identificados corretamente e os mapeamentos identificados (Fórmula 6).

$$Precisão = \frac{\#MapeamentosIdentificadosCorretamente}{\#MapeamentosIdentificados} \tag{6}$$

Cobertura é definida como a relação entre os mapeamentos identificados corretamente e os mapeamentos esperados pela versão oficial (Fórmula 7).

$$Cobertura = \frac{\#MapeamentosIdentificadosCorretamente}{\#MapeamentosCorretos} \tag{7}$$

F-Measure é a média harmônica da precisão e cobertura (Fórmula 8).

$$F - Measure = \frac{2 \times Precisão \times Cobertura}{Precisão + Cobertura} \tag{8}$$

A Tabela 5 (SNOMED-CT e NCI Thesaurus), a Tabela 6 (SNOMED-CT e ICD-9) e a Tabela 7 (MeSH e ICD-10) apresentam os resultados obtidos em termos de precisão, cobertura e f-measure ao aplicar o Algoritmo 1 no conjunto de dados estudado.

Resultados presentes na Tabela 5 indicam um decréscimo na precisão e f-measure para limiar igual a 0.5 quando o nível na ontologia fonte aumenta. O resultado melhora em termos de precisão, cobertura e f-measure para outros limiares. Os melhores resultados foram encontrados quando aumentamos o nível do contexto na ontologia fonte.

Nível fonte	**Limiar**	**Nível alvo**	**Precisão**	**Cobertura**	**F-Measure**
1	0.5	0	0.009	0.018	0.012
		1	0.042	0.101	0.060
		2	0.048	0.120	0.068
		3	0.048	0.127	0.069
	0.75	0	0	0	0
		1	0.088	0.082	0.085
		2	0.086	0.082	0.0841
		3	0.101	0.108	0.104
	0.9	0	0	0	0
		1	0.344	0.070	0.116
		2	0.378	0.089	0.144
		3	0.341	0.089	0.140
2	0.5	0	0.006	0.025	0.010
		1	0.031	0.139	0.051
		2	0.035	0.171	0.058
		3	0.034	0.184	0.058
	0.75	0	0.006	0.006	0.006
		1	0.089	0.120	0.102
		2	0.105	0.152	0.124
		3	0.108	0.165	0.130
	0.9	0	0.071	0.006	0.012
		1	0.357	0.095	0.012
		2	0.423	0.127	0.195
		3	0.407	0.139	**0.208**
3	0.5	0	0.004	0.019	0.006
		1	0.023	0.139	0.039
		2	0.029	0.190	0.050
		3	0.028	0.190	0.048
	0.75	0	0.005	0.006	0.006
		1	0.08	0.127	0.097
		2	0.104	0.177	0.125
		3	0.097	0.177	0.125
	0.9	0	0.071	0.006	0.012
		1	0.356	0.101	0.158
		2	0.434	0.146	**0.218**
		3	0.418	0.146	**0.216**

Tabela 5: Resultados dos mapeamentos derivados do SNOMED-CT e NCI

Os resultados apresentados na Tabela 6 (sobre os mapeamentos entre SNOMED-CT e ICD-9) são diferentes dos obtidos entre SNOMED-CT e NCI. Observamos um aumento na cobertura quando aumentamos o nível na ontologia fonte, mas a precisão diminui. A Tabela 6 apresenta os melhores resultados para os primeiros níveis na ontologia fonte e com menor limiar. Não foi possível observar grandes mudanças nos resultados ao aumentar o nível do contexto na ontologia alvo.

A Tabela 7 apresenta os resultados do refinamento para MeSH e ICD-10. Observamos uma melhora geral dos resultados quando aumentamos o nível do contexto na

Nível fonte	limiar	Nìvel alvo	Precisão	Cobertura	F-Measure
1	0.5	0	0.535	0.186	**0.276**
		1	0.340	0.163	**0.220**
		2	0.310	0.152	**0.204**
		3	0.296	0.145	0.196
	0.75	0	0.630	0.037	0.069
		1	0.461	0.044	0.081
		2	0.449	0.041	0.075
		3	0.439	0.041	0.075
	0.9	0	0.778	0.004	0.009
		1	0.692	0.006	0.011
		2	0.727	0.005	0.10
		3	0.75	0.006	0.011
2	0.5	0	0.325	0.181	0.233
		1	0.233	0.159	0.189
		2	0.241	0.167	0.198
		3	0.230	0.160	0.230
	0.75	0	0.487	0.046	0.084
		1	0.349	0.0455	0.080
		2	0.449	0.042	0.076
		3	0.382	0.048	0.085
	0.9	0	0.8	0.005	0.010
		1	0.687	0.007	0.014
		2	0.615	0.005	0.010
		3	0.615	0.005	0.010
3	0.5	0	0.256	0.177	0.209
		1	0.190	0.158	0.166
		2	0.200	0.159	0.177
		3	0.199	0.157	0.175
	0.75	0	0.444	0.051	0.091
		1	0.342	0.049	0.085
		2	0.360	0.050	0.089
		3	0.348	0.049	0.085
	0.9	0	0.833	0.006	0.013
		1	0.687	0.007	0.014
		2	0.687	0.007	0.014
		3	0.687	0.007	0.014

Tabela 6: Resultados dos mapeamentos derivados do SNOMED-CT e ICD-9

ontologia fonte.

Avaliamos comparativamente nossa proposta, considerando a vizinhança para obter os novos mapeamentos associados aos novos conceitos (Algoritmo 1), com a abordagem de computar a correspondência com toda a ontologia alvo. Aplicamos a abordagem sem contexto nos dados utilizando o limiar τ com o melhor resultado obtido para cada conjunto de dados no algoritmo 1.

A Tabela 8 apresenta os resultados de precisão, cobertura e f-measure obtidos para cada conjunto de dados utilizando a correspondência com todos os conceitos da ontologia alvo. Ao comparar os resultados, observamos que para o SCT-NCI, os resultados utilizando todos os conceitos do alvo teve melhores resultados. Para o SCT-ICD9, a abordagem com contexto é melhor. Para o MeSH-ICD10, as duas abordagens obtiveram resultados similares. Apesar disso, precisamos considerar o fato

de que aplicar a correspondência para toda a ontologia alvo tem uma complexidade de tempo pior que a abordagem com contexto.

Nível fonte	Limiar	Nível fonte	Precisão	Cobertura	F-Measure
1	0.5	0	0.059	0.048	0.053
		1	0.059	0.048	0.053
		2	0.059	0.048	0.053
		3	0.0625	0.048	0.054
	0.75	0	0.250	0.048	0.08
		1	0.250	0.048	0.08
		2	0.250	0.048	0.08
		3	0.250	0.048	0.08
	0.9	0	0	0	0
		1	0	0	0
		2	0	0	0
		3	0	0	0
2	0.5	0	0.067	0.095	0.078
		1	0.079	0.143	0.102
		2	0.0714	0.143	0.095
		3	0.0714	0.143	0.095
	0.75	0	0.250	0.048	0.08
		1	0.429	0.143	**0.214**
		2	0.429	0.143	**0.214**
		3	0.429	0.143	**0.214**
	0.9	0	0	0	0
		1	1.000	0.095	0.174
		2	1.000	0.095	0.174
		3	1.000	0.095	0.174
3	0.5	0	0.036	0.095	0.052
		1	0.044	0.143	0.067
		2	0.043	0.143	0.067
		3	0.043	0.143	0.066
	0.75	0	0.125	0.048	0.069
		1	0.333	0.143	0.2
		2	0.333	0.143	0.2
		3	0.333	0.143	0.2
	0.9	0	0	0	0
		1	1.000	0.095	0.174
		2	1.000	0.095	0.174
		3	1.000	0.095	0.174

Tabela 7: Resultados dos mapeamentos derivados do MeSH e ICD-10

Conjunto de dados	Limiar	Precisão	Cobertura	F-Measure
SCT-NCI	0.9	0.593	0.525	**0.557**
SCT-ICD9	0.5	0.264	0.042	**0.072**
MeSH-ICD10	0.75	0.312	0.238	**0.270**

Tabela 8: Resultado dos mapeamentos derivados da correspondência com todo os conceitos da ontologia alvo

6 Discussão

Esta pesquisa focou na criação de mapeamentos para atualizar o alinhamento entre ontologias baseado nos novos conceitos adicionados em novas versões das ontologias. Nossa abordagem tem três variáveis que afetam a qualidade do mapeamento: limiar, nível na ontologia alvo e nível na ontologia fonte. O limiar aumenta a precisão, mas diminui a cobertura. Isso é causado pelo fato de um maior limiar remover mapeamentos falsos positivos, mas remove mapeamentos corretos. Para dois conjuntos de dados (SCT-NCI e MeSH-ICD10) o aumenta na precisão compensou a queda na cobertura. Por outro lado, observamos para um conjunto de dados (SCT-ICD9) que um limiar alto tem efeitos negativos.

O nível na ontologia alvo aumenta a quantidade de conceitos candidatos para mapear ao aumentar o contexto na ontologia alvo. Isso significa que cada novo conceito adicionado tem mais opções para comparar. O aumento nos conceitos candidatos significa que tem chances maiores de encontrar o mapeamento correto, mas isso pode aumentar a chance de encontrar mapeamentos errados quando um conceito errado tem resultados melhores em termos da similaridade utilizada. Para dois conjuntos de dados (SCT-NCI e MeSH-ICD10), a precisão aumentou entre nível alvo 0 e 1 e a cobertura aumentou quando o nível alvo foi incrementado. Para um conjunto de dados (SCT-ICD9), precisão caiu entre o nível alvo 0 e 1 e a cobertura teve apenas pequenas variações ao alterar o nível alvo. Descobrimos que os resultados são dependentes das características do conjunto de dados.

O nível na ontologia fonte aumenta o contexto para encontrar conceitos candidatos da ontologia fonte. Nossa abordagem depende da existência de um conceito com mapeamento na versão anterior dentro do contexto do novo conceito. Se o nível fonte é baixo, os novos conceitos tem uma chance menor de encontrar um conceito mapeado na versão anterior. No pior caso, se não houver nenhum conceito mapeado na versão anterior, o novo conceito não é analisado para encontrar um novo mapeamento. Consequentemente, nesta abordagem, a derivação de mapeamentos relacionados com os novos conceitos dependem diretamente do nível da fonte. Encontramos resultados melhores ao aumentar o nível da fonte.

A análise dos resultados obtidos pela abordagem sem contexto indica que SCT-NCI teve resultados melhores usando tal abordagem. Por outro lado, a precisão teve resultados similares. SCT-ICD9 apresentou resultados melhores utilizando a abordagem contextual. Neste caso, a precisão teve bons valores utilizando a abordagem sem contexto, mas o f-measure sofreu com a baixa cobertura. No conjunto de dados MeSH-ICD10 os resultados foram similares para ambas as abordagens. Em resumo, a abordagem contextual implica em melhor precisão, mas a abordagem sem contexto obtêm uma cobertura melhor.

7 Conclusão

O mapeamento entre ontologias tem um papel central na integração semântica de dados semânticos. Porém, as atualizações no domínio do conhecimento aumenta a quantidade de novos conceitos na ontologia. Isso requer a manutenção dos conjuntos de mapeamento em tempo hábil de acordo com a dinâmica do conhecimento. Propusemos uma técnica que refina os alinhamentos existentes entre ontologias baseadas na evolução das ontologias. O algoritmo proposto considera o contexto dos conceitos em ambas as ontologias como um meio de encontrar as correspondências entre conceitos. A avaliação experimental com ontologias das ciências da vida alinhadas demonstraram a efetividade desta abordagem. Futuros trabalhos envolvem investigar outras heurísticas para atualizar os tipos de relações semânticas no processo de refinamento.

Agradecimentos

Agradecemos à Universidade Estadual de Campinas (UNICAMP) pela bolsa de iniciação científica concedida através do Programa Institucional de Bolsas de Iniciação Científica - PIBIC. Estendemos nossos agradecimentos a Fundação de amparo à pesquisa do estado de São Paulo (FAPESP) (projeto #2017/02325-5)[6].

Referências

[1] Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung Kwak. Snomed ct standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making*, 18, 12 2018.

[2] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, 2007.

[3] A. Groß, J. C. Dos Reis, M. Hartung, C. Pruski, and E. Rahm. Semi-Automatic Adaptation of Mappings between Life Science Ontologies. In *Proceedings The 9th International Conference on Data Integration in the Life Sciences*, pages 90–104, 2013.

[4] T. R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.

[5] Fayçal Hamdi, Brigitte Safar, Nobal B Niraula, and Chantal Reynaud. Taxomap alignment and refinement modules: Results for oaei 2010. In *Proceedings of the*

[6] As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão da FAPESP.

5th International Workshop on Ontology Matching, volume 689, pages 212–219, 2010.

[6] M. Hartung, A. Gross, and E. Rahm. COnto-Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. *Biomedical Informatics*, 46:15–32, 2013.

[7] Grzegorz Kondrak. N-gram similarity and distance. In *SPIRE*, 2005.

[8] Natalya F. Noy and Mark A. Musen. Anchor-prompt: Using non-local context for semantic matching. In *Workshop on ontologies and information sharing*, pages 63–70, 2001.

[9] Lorena Otero-Cerdeira, Francisco J. Rodríguez-Martínez, and Alma Gómez-Rodríguez. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949 – 971, 2015.

[10] C. Pruski, J. C. Dos Reis, and M. Da Silveira. Capturing the relationship between evolving biomedical concepts via background knowledge. In *Proceedings of the 9th International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS'16)*, 2016.

[11] Julio Cesar Dos Reis, Cédric Pruski, Marcos Da Silveira, and Chantal Reynaud-Delaitre. Dykosmap: A framework for mapping adaptation between biomedical knowledge organization systems. *Journal of Biomedical Informatics*, 55:153 – 173, 2015.

[12] Md. Hanif Seddiqui and Masaki Aono. Anchor-flood: Results for oaei 2009. In *Proceedings of the 4th International Conference on Ontology Matching - Volume 551*, OM'09, pages 127–134, Aachen, Germany, Germany, 2009. CEUR-WS.org.

[13] Yoones A Sekhavat and Jeffrey Parsons. Sesm: Semantic enrichment of schema mappings. In *Proceedings of the 29th International Conference on Data Engineering Workshops (ICDEW 2013)*, pages 7–12. IEEE, 2013.

[14] Pavel Shvaiko and Jérôme Euzenat. Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176, 2013.

[15] Suzette Kruger Stoutenburg. Acquiring advanced properties in ontology mapping. In *Proceedings of the 2nd PhD Workshop on Information and Knowledge Management (PIKM 2008)*, pages 9–16. ACM, 2008.

[16] Songmao Zhang and Olivier Bodenreider. Experience in aligning anatomical ontologies. *International journal on Semantic Web and information systems*, 3(2):1, 2007.