Alinhamento de grandes Ontologias com recurso de banco de dados *NoSQL* e utilização de *workflow* científico

Luciana de Sá Silva Perciliano, Fernanda Araujo Baião Amorim (orientadora)

Departamento de Informática Aplicada
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur, 458 – Rio de Janeiro– RJ – Brasil

{luciana.perciliano@uniriotec.br, fernanda.baiao@uniriotec.br}

Nível: Mestrado

Ingresso no Programa: 03/2017

Previsão de Defesa: 04/2019

Etapas concluídas: Definição do Problema, Proposta de Pesquisa apresentada em seminário na UNIRIO.

Publicações: Alinhamento de Ontologias com Suporte de um Sistema Gerenciador de *Workflows* Científicos – ERAD (IV Escola Regional de Alto Desempenho do Rio de Janeiro) – 05/2018.

Resumo. *Uma ontologia é um artefato que em um mesmo domínio pode ser representado de diferentes formas em sistemas distintos. O alinhamento de Ontologias é uma técnica que visa resolver o problema da heterogeneidade semântica entre esses sistemas. O alinhamento de grandes Ontologias (milhões ou bilhões de correspondências possíveis entre duas entidades de diferentes ontologias) é um desafio na área e demanda bastante tempo de execução e recursos computacionais. Esse desafio, pode ser verificado na iniciativa anual OAEI (Ontology Alignment Evaluation Initiative) que objetiva a comparação e avaliação dos sistemas de alinhamento. Dado que um banco de dados NoSQL é eficiente no armazenamento de grande volume de dados e que através de um Sistema de Gerência de Workflows Científicos (SGWfC) é possível a execução paralela e a execução em ambiente de nuvem é escalável, o objetivo da proposta de dissertação consiste em verificar se a utilização de desses recursos, resultam em menor tempo na execução e na escalabilidade em um processo de um sistema de alinhamento de Ontologias.*

1. Introdução

Uma ontologia, sob o viés computacional, é reconhecida como um artefato capaz de representar formalmente uma conceituação compartilhada [Gruber, 1993]. No entanto, existem diferentes formas de representar uma conceituação sobre um mesmo domínio, o que dificulta a troca de informações e entendimento entre diferentes sistemas, conhecido como problema da heterogeneidade semântica na ontologia. Uma das soluções para auxiliar esse problema é a aplicação de técnicas de alinhamento de Ontologias.

O alinhamento de Ontologias é um tópico que vem sendo pesquisado há algum tempo, com resultados positivos de bastante impacto e alguns desafios ainda em aberto. Dois desafios no alinhamento de Ontologias citados por [Shvaiko e Euzenat, 2013] são a eficiência e a escalabilidade. Os sistemas de alinhamento de Ontologias em sua maioria usam o armazenamento em memória e, com a existência cada vez mais frequente de ontologias de grande tamanho, há uma demanda bastante expressiva no consumo de memória, e o tempo requerido para o processamento das medidas de similaridades e técnicas correlatas, o que pode acarretar o elevado tempo de execução para seu término. A proposta de dissertação é apresentar uma estratégia para executar em paralelo as etapas de particionamento e de alinhamento de grandes ontologias baseada em Sistema de Gerência de *Workflows* Científicos (SGWfC) e Sistema Gerenciador de Banco de Dados (SGBD) *NoSQL* com o objetivo de aumentar o desempenho e a escalabilidade na execução de sistemas de alinhamento de Ontologias. Em particular, será aplicado o SGWfC SciCumulus [Silva et al. 2014] para execução paralela e o SGBD *NoSQL* Neo4J como recurso no processo de alinhamento de Ontologias.

Esse documento utiliza a seguinte disposição: Na seção 2 é realizado a Apresentação do Problema. A seção 3 descreve a Fundamentação Teórica. Nas seções 4 e 5 é descrito a Proposta da Solução e o Projeto de Avaliação da Solução. Na seção 6 os Trabalhos Relacionados. Na seção 7 é apresentada as Publicações. E na seção 8 a Conclusão.

2. Apresentação do Problema

A OAEI (*Ontology Alignment Evaluation Initiative*) é uma iniciativa anual e internacional que avalia e compara o desempenho dos sistemas de alinhamentos de Ontologias. A Figura 1 demonstra os *datasets* e os sistemas que participaram em 2017.

System	ALIN	AML	CroLOM	DiSMatch-ar	DiSMatch-sg	DiSMatch-tr	I-Match	KEPLER	Legato	LogMap	LogMap-Bio	LogMapLt	njuLink	ONTMAT	POMap	RADON	SANOM	Silk	WikiV2	XMap	YAM-BIO	Total=21
Confidence	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	-	✓	-	✓	✓	✓	✓	-	✓	-	16
anatomy	●	●	○	○	○	○	○	●	○	●	●	●	○	○	●	○	●	○	●	●	●	11
conference	●	●	○	○	○	○	○	●	○	●	○	●	○	●	●	○	●	○	●	●	○	10
largebio	○	●	○	○	○	○	○	◐	○	●	●	●	○	○	◐	○	◐	○	◐	●	●	10
phenotype	○	●	○	●	●	●	○	◐	○	●	●	◐	○	○	◐	○	○	○	○	◐	◐	11
multifarm	○	●	●	○	○	○	○	●	○	●	○	○	○	○	○	○	◐	○	●	◐	○	7
interactive	●	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	●	○	4
process model	○	●	○	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	3
instance	○	●	○	○	○	○	●	○	●	●	○	○	◐	○	○	○	○	○	○	○	○	5
hobbit ld	○	●	○	○	○	○	●	○	○	○	○	○	○	○	○	◐	○	◐	○	○	○	4
total	3	9	1	1	1	1	3	5	1	8	3	4	1	1	4	1	4	1	4	6	3	65

Figura 1. Resultado OAEI 2017 [Achichi et al. 2017]

As bolas pretas indicam que o sistema conseguiu executar com sucesso o alinhamento, as brancas significam que não foi executado e as metade preta e metade branca significam que não foi possível a conclusão da execução em um determinado *dataset* [Achichi et al. 2017]. De acordo com [Achichi et al. 2017]: "Não houve progresso significativo no que diz respeito à capacidade de sistemas de alinhamento para lidar com grandes ontologias e *datasets*, seja na correspondência de ontologia tradicional ou na correspondência de instâncias." e "Não houve melhora notável em relação aos tempos de execução do sistema". Os *datasets Disease* e *Phenotype Track* (*phenotype*) e *Large BioMed Track* (*largebio*) possuem ontologias de grande tamanho [Achichi et Al. 2017] que contêm respectivamente 138.143.706 e 1.224.924.624 correspondências máximas possíveis entre ontologias que compõe esses *datasets*. Alguns sistemas como: POMAP, SANOM, KEPLER e Wiki2, não terminaram a execução no tempo máximo atribuído pela iniciativa de quatro horas o alinhamento entre Ontologias do *dataset largebio* [Achichi et al. 2017]. Sendo assim, o problema do alinhamento de grandes Ontologias ainda é um desafio na área, conforme foi descrito em [Shvaiko e Euzenat, 2013].

3. Fundamentação Teórica

O alinhamento de Ontologias é o processo que, a partir de um par de ontologias, visa encontrar automaticamente o subconjunto de correspondências semânticas entre pares de entidades (classes, atributos, relacionamentos ou instâncias) das ontologias de entrada. Esse pode ser realizado de forma manual, semi-automática ou em tempo de execução. A Figura 2(a) ilustra um exemplo de alinhamento de Ontologias, no qual as classes estão colocadas dentro dos retângulos de cantos arredondados. As setas verticais exibem o relacionamento especialização-generalização, do mais específico "Literatura" para o mais genérico "Monografia". Os atributos estão colocados logo após as setas tracejadas. Nesse exemplo, as setas azuis representam as correspondências entre as duas ontologias (O1-Produto e O2-Monografia) [Shvaiko e Euzenat, 2013] [da Silva et. al. 2016]. Formalmente, uma correspondência é representada por um par de entidades e o tipo de relação existente entre elas pode ser: equivalência (=), disjunção (⊥) ou generalização (⊒). O alinhamento entre O1 e O2 é o conjunto das correspondências encontradas [Lopes, 2014]. No processo do alinhamento de Ontologias ilustrado na Figura 2(b), através das ontologias de entrada O1 e O2 é gerado o alinhamento A'. Esse pode ter passado por um outro processo de alinhamento A anteriormente.

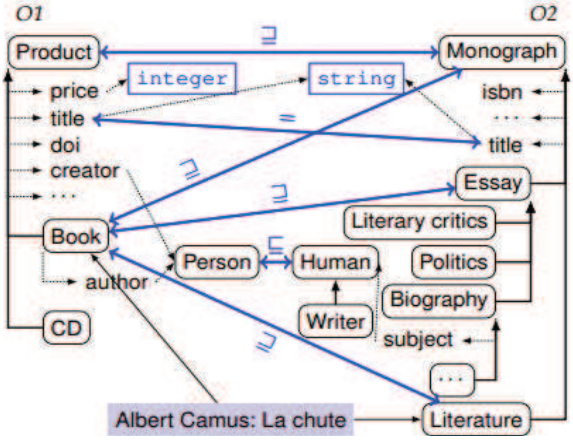

Figura 2(a). Alinhamento entre duas Ontologias [Shvaiko e Euzenat, 2013]

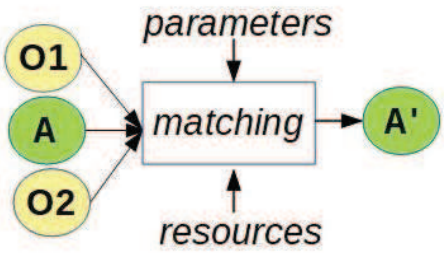

Figura 2(b). Processo de Alinhamento e Ontologias [Shvaiko e Euzenat, 2013]

Os parâmetros podem ser pesos e *thresholds*, e os recursos externos uma base léxica como a *Wordnet*, ontologias de referência ou relatórios de padrões e anti-padrões de alinhamento [Lopes, 2014], [da Silva et. al. 2016]. O sistema de alinhamento de Ontologias denominado ALIN [da Silva et al. 2016] busca aumentar a eficiência da participação do especialista (usuário conhecedor da ontologia que está sendo alinhada) e utiliza anti-padrões (combinação de correspondências que geram inconsistências) de alinhamento para melhorar a qualidade do alinhamento final. A utilização dessas características aumenta tanto a precisão como a cobertura do alinhamento obtido se comparado a abordagem que não as utiliza [da Silva et al. 2016]. A arquitetura do ALIN [da Silva et. al. 2016], utiliza os seguintes itens: APIs em Java (*Weka*, com rotinas estatísticas e de KDD), *Simmetrics* (métricas de similaridade baseadas em *strings*), WS4J (métricas linguísticas baseadas na *Wordnet*) e *Alignment* (possui rotinas para manipulação de ontologias escritas em OWL).

Nos bancos de dados *NoSQL* orientados a grafos, os dados são armazenados na forma de um grafo e servem para armazenamento de grandes volumes de dados e sua estrutura se assemelha ao esquema de uma ontologia. O banco de dados orientado a grafos Neo4J, possui uma estrutura requerida para armazenamento de ontologias [Pramanik, 2016]. No Neo4J, todas as entidades das ontologias (classes, propriedade de dados e propriedades de objetos) podem ser representadas pelos nós (vértices) e seus relacionamentos preservados através das arestas, assim como novos relacionamentos entre entidades de diferentes ontologias para representar correspondências no processo do alinhamento, com suas propriedades. A Figura 3, exemplifica relacionamentos entre entidades de duas diferentes ontologias do *dataset Conference* [Achichi et al. 2017]. Os nós representam as entidades, as arestas os relacionamentos, sendo que o relacionamento em destaque possui propriedades, por exemplo, NOME_DA_METRICA: “LIN” e VALOR_DA_METRICA: “0.16966131697611691” armazenadas no Neo4J.

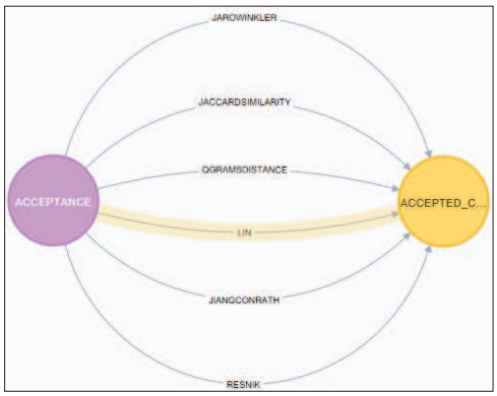

Figura 3. Exemplo de Ontologias no Neo4J

O processo do alinhamento de Ontologias no ALIN possui várias atividades, como na geração do conjunto de correspondências candidatas e na classificação e modificação do conjunto de correspondências candidatas que podem utilizar recursos, como o armazenamento em banco de dados. Esse processo se assemelha a um *workflow* científico que é uma forma abstrata de representar um experimento científico como uma sequência de atividades, sendo que o seu fluxo de dados pode sofrer diversas variações, incluindo a aplicação de parâmetros diferentes, conjuntos de dados de entrada distintos, algoritmos alternativos para a mesma tarefa, entre outros [de Oliveira et al. 2010]. Os Sistemas de Gerência de *Workflows* Científicos (SGWfC) executam e gerenciam os

fluxos de dados de um *workflow* científico [de Oliveira et al. 2010]. De acordo com [Silva & Mattoso, 2014]: "A gerência da dependência do fluxo de dados do *workflow* é um dos diferenciais dos SGWfC em relação a soluções que programam esse controle por *scripts* ou Hadoop, ou de forma independente (manual)". Diferentemente do conceito de MapReduce em que é necessário para os cientistas programarem a execução de experimento científico [de Oliveira et al. 2010], o SciCumulus é um *middleware* que promove uma execução de um *workflow* em um ambiente de nuvem, sendo possível controlar e monitorar as execuções das atividades desse, de forma a isolar o usuário da complexidade de um ambiente de nuvem e da distribuição dos dados. Além disso, um benefício no uso do SciCumulus é o banco de proveniência que permite aos usuários a consulta dos dados históricos, análise e monitoração do fluxo de dados em todo seu ciclo de vida inclusive em tempo de execução [de Oliveira et al. 2010] [Silva & Mattoso, 2014]. Diversos SGWfC existentes na literatura têm suporte para acesso intensivo a dados aplicando técnicas de processamento de alto desempenho, incluindo suporte a paralelismo e distribuição [Liu et al. 2015], por exemplo, Pegasus e Swift, nesses SGWfC os dados de proveniência são disponibilizados apenas no final da execução do *workflow* [Silva & Mattoso, 2014].

4. Proposta de Solução

A proposta de solução é utilizar um SGWfC em particular o SciCumulus para executar, monitorar e controlar a execução de um sistema de alinhamento de Ontologias em um ambiente de nuvem de forma paralela, pelos motivos descritos na seção 3. E a utilização do recurso de armazenamento de dados de ontologias no Neo4J (conforme modelagem descrita na seção 3) no processo de alinhamento de Ontologias. A contribuição principal desta proposta é uma estratégia de alinhamento de grandes ontologias que executa em paralelo as etapas de particionamento e de *matching* das ontologias. A figura 4 ilustra a arquitetura da solução. Dado 2 ontologias O1 e O2, elas passam por um processo de "*partition of ontologies*" em que serão particionadas, ou seja, ao invés de todos os pares das entidades de O1 serem alinhados com os de O2, eles serão divididos em vários blocos. Depois dessa divisão, por exemplo, O1' e O2', passa pelo processo de "*matching*" que recebe "*parameters*" e em que "*resources*" podem ser consultados como um banco *NoSQL* sempre que necessário, essa etapa gera um alinhamento parcial A'. Isso ocorre até que todos os pares de subontologias de O1 e O2 gerem o alinhamento parcial A'. Quando todos os alinhamentos parciais A' estiverem finalizados, esses passam pelo "*combine matching*" em que todos os alinhamentos A' serão combinados, gerando um único alinhamento final das correspondências entre O1 e O2 denominado AF.

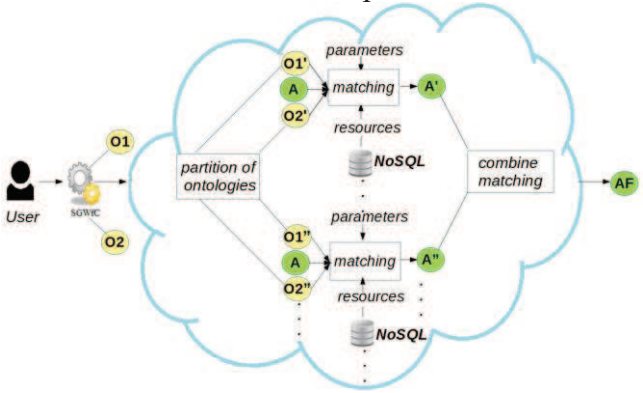

Figura 4. Arquitetura da Solução

5. Projeto de Avaliação da Solução

Utilizar o SGWfC SciCumulus [Silva et al. 2014] para gerenciar a execução paralela em um ambiente de nuvem do sistema de alinhamento de Ontologias ALIN [da Silva et al. 2016] com os *datasets phenotype* e *largebio* disponíveis na página do OAEI 2017[1], com a utilização de recurso no processo de alinhamento o banco de dados orientado a grafos Neo4J. A precisão, cobertura e medida-F são formas de avaliar a qualidade do alinhamento de Ontologias que foi gerado [da Silva et al. 2016]. Os resultados serão avaliados com relação ao tempo de execução e também com relação à precisão, cobertura e medida-F (de forma a garantir que a execução em paralelo não impactou a qualidade dos resultados encontrados) comparando-os com os resultados do OAEI 2017.

6. Trabalhos Relacionados

Existem alguns trabalhos que utilizam a execução paralela em nuvem para diminuir o tempo de execução do alinhamento de Ontologias, como em [Araújo et al. 2015] que utiliza o particionamento das ontologias em subontologias e a abordagem MapReduce no processo do alinhamento das Ontologias, um dos seus trabalhos futuro é investigar como grandes ontologias podem ser particionadas em paralelo com algoritmo de particionamento PAP (*Partition, Anchor, Partition*). Em [Araújo et al. 2016] é proposto uma técnica de *load balancing* e a abordagem MapReduce para o alinhamento de grandes ontologias e possui como um dos seus trabalhos futuro a pesquisa de como particionar em paralelo grandes ontologias.

7. Publicações

As tabelas 1 e 2 apresentam respectivamente as principais publicações relacionadas e o plano de publicação.

Tabela 1. Principais publicações relacionadas

Publicação
Achichi, M., Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., ... & Zamazal, O. (2017*). Results of the Ontology Alignment Evaluation Initiative 2017*. In Proceedings of the 12th International Workshop on Ontology Matching-Volume 2032 (pp. 61-113)
Araújo, T. B., Pires, C. E., da Nobrega, T. P., & Nascimento, D. C. (2015). *A parallel approach for matching large-scale ontologies*. Journal of Information and Data Management, 6(1), 18.
da Silva, J., Baião, F. A., & Revoredo, K. (2016). Alinhamento Interativo de Ontologias usando Anti-Padrões de Alinhamento: Um primeiro Experimento. XII Brazylian Symposiumon Information Systems, Florianopolis, SC, p. 208-215.
de Oliveira, D., Ogasawara, E., Baião, F., & Mattoso, M.(2010). Scicumulus: *A light-weight cloud middleware to explore many task computing paradigm in scientific workflows*. IEEE 3rd International Conference on(pp. 378-385).
Shvaiko, P., & Euzenat, J. 2013. *Ontology matching: state of the art and future challenges*.IEEE Transactions on Knowledge and Data Engineering, 25(1), p. 158-176.

Tabela 2. Plano de publicação

Evento/ Periódico	Data prevista de submissão	Conteúdo do trabalho	Situação
SBBD-2019	05/2019	Resultados alcançados comparando com resultados do OAEI 2018	não-submetida
Semantic Web journal	12/2018	Resultados do alinhamento de grandes Ontologias com SGWfC e banco de dados *NoSQL*	não-submetida

[1] *http://oaei.ontologymatching.org/2017/*

8. Conclusão

O alinhamento de grandes Ontologias ainda é um desafio, conforme [Achichi et al. 2017]. A proposta de dissertação propõe uma estratégia que executa em paralelo as etapas de particionamento e de alinhamento de grandes Ontologias e utiliza o SGWfC SciCumulus para que esse ocorra em um ambiente de nuvem de forma paralela e também o banco de dados *NoSQL* Neo4J como recurso no processo de alinhamento, para comprovação da hipótese a ser testada. Presume-se que as características dessas tecnologias podem contribuir para a diminuição do tempo de execução e na escalabilidade de um sistema de alinhamento de grandes Ontologias.

Referências

Achichi, M., Cheatham, M., Dragisic, Z., Euzenat, J., Faria, D., Ferrara, A., ... & Zamazal, O. (2017). *Results of the Ontology Alignment Evaluation Initiative 2017*. In Proceedings of the 12th International Workshop on Ontology Matching-Volume 2032 (pp. 61-113)

Araújo, T. B., Pires, C. E., da Nobrega, T. P., & Nascimento, D. C. (2015). *A parallel approach for matching large-scale ontologies*. Journal of Information and Data Management, 6(1), 18.

Araújo, T. B., Pires, C. E. S., da Nóbrega, T. P., & Nascimento, D. C. (2016). *A fine-grained load balancing technique for improving partition-parallel-based ontology matching approaches*. Knowledge-Based Systems, 111, 17-26.

da Silva, J., Baião, F. A., & Revoredo, K. (2016). Alinhamento Interativo de Ontologias usando Anti-Padrões de Alinhamento: Um primeiro Experimento. XII Brazylian Symposium on Information Systems, Florianopolis, SC, p. 208-215.

de Oliveira, D., Ogasawara, E., Baião, F., & Mattoso, M., 2010. Scicumulus: *A light-weight cloud middleware to explore many task computing paradigm in scientific workflows*. In Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on (pp. 378-385). IEEE.

Gruber, T. R. (1993). *A translation approach to portable ontology specifications*. Knowledge acquisition, 5(2), 199-220

Liu, J., Pacitti, E., Valduriez, P., & Mattoso, M. (2015). *A survey of data-intensive scientific workflow management*. Journal of Grid Computing, 13(4), 457-493.

Lopes, V., 2014. Alinhamento Interativo de Ontologias: Uma Abordagem Baseada em *Query-by-Committee* (Doctoral dissertation, Msc Dissertation, Rio de Janeiro, Rj, Brazil: Universidade do Estado do Rio de Janeiro (UNIRIO)).

Pramanik, Gopal. (2016). *Ontology and Neo4J Graph Database*. Scientific Voyage, Volume -2, Issue - 2, Date Of Publication - May, 2016.

Shvaiko, P., & Euzenat, J. (2013*). Ontology matching: state of the art and future chal-lenges*. IEEE Transactions on Knowledge and Data Engineering, 25(1), p. 158-176

Silva, V., Oliveira, D., & Mattoso, M. (2014). SciCumulus 2.0: Um Sistema de Gerência de *Workflows* Científicos para Nuvens Orientado a Fluxo de Dados. Sessão de Demos do XXIX Simpósio Brasileiro de Banco de Dados.