1
Introdução

O progresso das tecnologias de informação e comunicação tem disponibilizado uma enorme quantidade de informações, uma vez que a quantidade de fontes de informação tem aumentado significativamente (Shvaiko & Euzenat, 2008). Hoje a Web garante, de maneira prática, que grande parte desses recursos estejam disponíveis, sendo considerada atualmente o maior repositório de informações existente, com um número de usuários cada vez maior e em franco processo de evolução, se tornando uma enorme base de conhecimento.

Contudo, com o crescimento dos recursos de informação disponíveis na web, surge o problema de gerenciar a heterogeneidade entre estes recursos. Várias soluções para lidar com a heterogeneidade foram propostas e, especificamente, lidar com a automatização da integração de fontes distribuídas de informação (Shvaiko & Euzenat, 2008). Entre essas soluções, as tecnologias semânticas tem atraído grande atenção, vide a pesquisa do Gartner Group que identificou as tecnologias semânticas como uma das 10 tecnologias disruptivas[1] para o quadriênio 2008–2012 (Gartner, 2008).

Porém, embora a Web seja uma poderosa fonte provedora de informação, a falta de confiabilidade e a pouca estruturação faz com que o montante de informação disponibilizada possua baixa qualidade para fins computacionais e, assim, alto valor informativo não aproveitado. Por esses motivos, embora os atuais mecanismos de busca existentes na WWW tenham evoluído de forma a manipular a informação existente com maior precisão de resposta, ainda assim não são capazes de suprir todas as necessidades dos desenvolvedores de aplicações que consomem essas informações, como sistemas multi-agentes ou sistemas de coleta de informações de um dado domínio na web. Devido ao excesso de informação disponibilizada, oriunda de diversas fontes, é possível absorver e empregar conceitos – oriundos destes domínios – de forma equivocada,

[1] Tecnologias ou inovações disruptivas é um termo criado por Clayton Christensen e Joseph Bower no artigo *Disruptive Technologies: Catching the Wave* (Bower & Christensen, 1995), a qual descreve a inovação tecnológica, produto, ou serviço, que utiliza uma estratégia "disruptiva", em vez de "revolucionária" ou "evolucionária", para derrubar uma tecnologia existente dominante no mercado.

pois, em muitos casos, depara-se com a falta de clareza e ambigüidade com que eles são tratados. Nesse contexto, identificamos a necessidade da construção de mecanismos que tornem a representação de dados e conceitos de modo mais formal, retirando a ambiguidade e permitindo um melhor processamento da informação. Uma das formas de se representar conceitos de domínios de maneira mais formal é através de ontologias.

Uma ontologia tipicamente fornece um vocabulário descrevendo um domínio de interesse e uma especificação do significado dos termos deste vocabulário (Euzenat & Shvaiko, 2007). Dependendo da precisão dessa especificação, a noção de ontologia engloba vários modelos de dados e de conceitos como, por exemplo, classificações, esquemas de bancos de dados e teorias axiomatizadas (Shvaiko & Euzenat, 2012). Segundo Fensel (2004), ontologias são vistas como uma solução para interoperabilidade em muitas aplicações, como integração de banco de dados, sistemas *peer-to-peer*, comércio eletrônico, serviços web e redes sociais.

A construção de uma única ontologia que englobe todas as áreas de conhecimento é extremamente trabalhosa. O projeto OpenCyc (Matuszek et al, 2006) pode ser um exemplo do quão trabalhoso pode ser a definição formal de conceitos e relações em uma ontologia. O projeto demorou quase 20 anos para formalizar a primeira versão de uma ontologia sobre conhecimento de senso comum e ainda assim recebeu diversas críticas, sobre o tratamento insatisfatório de alguns conceitos, documentação limitada, entre outras (Bertino et al, 2001, p.275). Além disso, a simples construção de ontologias de domínio, isoladamente, resolve alguns, mas não todos os problemas no que se refere a um maior enriquecimento semântico. Mesmo ontologias criadas através de metodologias e técnicas que garantam sua conformidade com os conceitos de compartilhamento e reuso que as regem (Souza et al, 2008), não são capazes de garantir uma visão consensual e nem de abranger todos os conceitos necessários por quaisquer aplicações. Ainda, ontologias não são estruturas inertes, ou seja, são passíveis de alterações e de adaptações, seja por adequação de modelagem ou por descoberta de novo conhecimento ou mudança em requisitos de negócio (Stojanovic et al, 2002).

Em sistemas abertos ou em evolução, como a web semântica, diferentes agentes podem adotar diferentes ontologias. Assim, segundo Euzenat & Shvaiko (2007), a simples utilização de ontologias não reduz a heterogeneidade, mas leva os problemas de heterogeneidade para um nível mais alto. Dessa forma, tornam-se necessários mecanismos que permitam a interoperabilidade semântica. Entende-se por interoperabilidade semântica a capacidade de dois ou mais

sistemas heterogêneos e distribuídos trabalharem em conjunto, compartilhando as informações entre eles com entendimento comum de seu significado (Buranarach, 2001). Várias propostas de tratamento para esse problema já foram divulgadas no meio científico (Felicissimo et al, 2005; Euzenat & Shvaiko, 2007; Lanzenberger & Sampson, 2007; Duong et al, 2008; Leme et al, 2009; Li et al, 2009; Leme et al, 2010; Shvaiko & Euzenat, 2012), embora ainda não exista uma solução final e de propósito geral (Shvaiko & Euzenat, 2012).

Entre as tecnologias semânticas utilizadas para compatibilizar ontologias, estão as técnicas de *ontology matching*, as quais traduzimos nesse trabalho como técnicas de *alinhamento de ontologias*. Alinhamento de ontologias é o processo de determinar correspondências entre entidades de dois ou mais modelos (Giunchiglia et al, 2007).

O processo de alinhamento é uma importante operação em aplicações tradicionais, como integração de modelos (Rahm & Bernstein, 2001; Do et al, 2003) ou *data warehouses* (Bernstein & Rahm, 2000; Do, 2007). Essas aplicações são caracterizadas por modelos com estruturas heterogenêas que são analisados e alinhados de forma manual ou semi-automática, geralmente durante o tempo de projeto. Em tais aplicações, o alinhamento é um pré-requisito para o funcionamento do sistema. Outros sistemas são caracterizados por sua dinâmica, como sistemas multi-agentes (Wiesman et al, 2002; Souza et al, 2009a,b; Santos et al, 2011), sistemas *peer-to-peer* (Zaihrayeu, 2006; Staab & Stuckenschmidt, 2006), serviços web (Sycara et al, 2003; Fensel et al, 2011) e outras aplicações que consomem dados da web, como as baseadas em dados ligados (Volz et al, 2009; Heath & Bizer, 2011; Souza et al, 2012a). Tais aplicações, ao contrário das aplicações tradicionais, requerem uma operação de alinhamento em tempo de execução.

Este trabalho trata do problema de alinhamento de ontologias, objetivando contribuir para as diversas aplicações que dependem dessa operação. Na seção 1.1 é apresentada a relevância desse trabalho, enquanto nas seções 1.2 e 1.3 são apresentados os objetivos do trabalho e a metodologia aplicada para solução do problema, respectivamente. Por fim, a seção 1.4 apresenta a organização desta tese de doutorado.

1.1 Motivação

Tratar a heterogeneidade não é uma questão simples. Por exemplo, a maioria dos relatórios de auto-avaliação da área de Banco de Dados reconhecem o problema que é lidar com a heterogeneidade semântica, isto é, de lidar

com variações no significado ou ambiguidade na interpretação de entidades, continua sendo um problema em aberto (Agrawal et al, 2008). O problema de alinhamento de ontologias não é um problema recente e, nos últimos anos, muitos esforços foram realizados para resolver esse problema. Verificando, por exemplo, o número de artigos publicados[2] sobre alinhamento de ontologias nas principais conferências e periódicos científicos entre 2002 e 2011 (figura 1.1), é possível observar que houve um crescimento no número de artigos publicados à partir de 2002 e, após 2008, uma desaceleração.

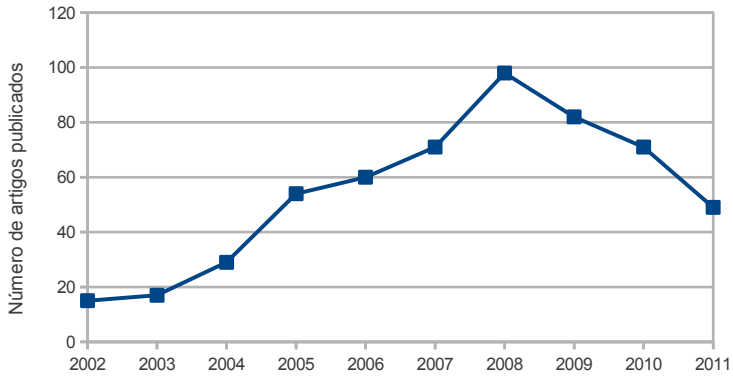

Figura 1.1: Quantidade de artigos publicados por ano

Essa desaceleração já foi identificada por Shvaiko & Euzenat (2012). Segundo os autores, as publicações dos últimos anos mostram que o campo de alinhamento de ontologias realizou uma grande melhoria, embora a velocidade com que as melhorias ocorrem esteja desacelerando. Ainda segundo os autores, era esperada a diminuição na quantidade de artigos publicados, uma vez que os diversos desafios conhecidos que ainda persistem se mostram mais difíceis de serem tratados, muitas vezes necessitando de soluções mais especializadas para um certo problema.

Do ponto de vista da adoção dessas tecnologias, o grupo Knowledge Web[3] publicou um relatório com resultados de uma pesquisa com vários pesquisadores e profissionais do mercado para avaliar a maturidade e expectativas com as tecnologias semânticas (Cuel et al, 2007). O grupo posicionou diferentes tecnologias semânticas em uma curva de expectativa[4] proposta pela Gartner. Em relação à tecnologia de alinhamento de ontologias,

[2]Dados retirados do site http://www.ontologymatching.org/publications, o qual registra as principais publicações sobre alinhamento de ontologia. Último acesso em 25/03/2012.

[3]http://knowledgeweb.semanticweb.org

[4]A curva de expectativa da Gartner, chamada de *Hype Cycle*, é uma representação gráfica da maturidade, adoção e aplicação social de tecnologias específicas. Ver http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

tanto pesquisadores quanto profissionais do mercado concordaram que essa tecnologia estava chegando na fase de expectativa com uma duração de 5 a 10 anos para adoção em larga escala. Ainda, o relatório aponta que muitos desafios ainda necessitam ser atacados antes que a tecnologia de alinhamento de ontologia seja completamente adotada em larga escala.

Um dos desafios identificados por Shvaiko & Euzenat (2008) no seu artigo *Ten Challenges for Ontology Matching* está a seleção de alinhadores e auto-configuração desses alinhadores. Como muitos alinhadores foram propostos na literatura, constata-se que vários alinhadores provêem bons resultados em alguns casos e resultados não tão bons em outros. Assim, fazer uso de diversos alinhadores e descobrir quais são melhores aplicados para cada caso é um desafio para o campo de alinhamento de ontologias. Somado a esse desafio, encontra-se a necessidade de reavaliar os alinhadores utilizados em ambientes dinâmicos, como a web. É natural que nesse tipo de ambiente, as características das aplicações estão constantemente em mudança. Assim, as abordagens que tentam calibrar e adaptar automaticamente soluções de alinhamento para as configurações nas quais uma aplicação opera são de grande importância. Algumas propostas foram realizadas para calibrar uma biblioteca de alinhadores de esquemas em tempo de projeto, como o trabalho de Lee et al (2007). Contudo, em ambientes dinâmicos, a calibragem em tempo de execução possui uma maior relevância.

Em 2012, Shvaiko & Euzenat (2012) revisitam seus desafios futuros para o campo de alinhamento de ontologia e constatam que (1) o problema de calibragem de alinhadores ainda se encontra em aberto e é um problema promissor, (2) o problema do desempenho dos alinhadores precisa ser tratado de forma mais séria.

Entre as abordagens atuais, está o meta-alinhamento de ontologias, ou seja, um *framework* para combinar um conjunto de alinhadores de ontologias escolhidos. Embora existam algumas propostas nessa direção, conforme será apresentado no capítulo 2, ainda existe a necessidade de abordagens que possam ser aplicadas em tempo de execução (ao invés de tempo de projeto). Para tal, essas soluções necessitam ter um bom desempenho. Meta-alinhadores tendem a degradar o tempo necessário para o alinhamento, uma vez que vários alinhadores precisam ser analisados e a busca pela melhor calibragem é custosa. Segundo Shvaiko & Euzenat (2012), nas avaliações de alinhadores realizadas na OAEI[5], constatou-se que as abordagens podem levar vários minutos para completar a tarefa de alinhamento, chegando até a horas ou

[5]Ontology Alignment Evaluation Initiative

mesmo dias em certas abordagens. De fato, conforme Lee et al (2007) analisam, abordagens de seleção e calibragem de alinhadores são geralmente custosas em relação ao tempo de processamento. Contudo, ainda segundo Lee et al (2007), a utilização dessas abordagens melhora significativamente a precisão do alinhamento. Associada a essa questão, está a dificuldade da realização dos alinhamentos em larga escala (Euzenat et al, 2011), ou seja, em esquemas com milhares de entidades, o que é o caso de fontes como Wordnet[6] e DBPedia[7].

1.2 Objetivos

Este trabalho tem como objetivo contribuir para o problema de alinhamento de ontologias. Mais especificamente, este trabalho objetiva tratar o problema de meta-alinhamento de ontologias e a calibragem dos alinhadores escolhidos sem detrimento do desempenho do sistema. Ou seja, almeja-se elaborar uma solução que possa ser aplicada em tempo de execução e em alinhamentos de larga escala.

Uma vez que meta-alinhadores se utilizam de alinhadores independentes, é necessária uma solução que possa utilizar distintos alinhadores e em quantidade variada. Uma solução que possa ser utilizada em tempo de execução deve ser uma solução que gere resultados praticamente tão rápido quanto o tempo de execução de cada alinhador, podendo ser utilizada durante o processo de alinhamento necessário em sistemas diversos, como serviços web, sistemas multi-agentes e outros.

1.3 Enfoque de solução e método de avaliação da pesquisa

Selecionar, dentre um conjunto de alinhadores, quais alinhadores fornecerão um melhor alinhamento entre duas ontologias não é um problema trivial. Os alinhadores que retornarão melhor valor na comparação entre duas entidades são dependentes das duas ontologias a serem analisadas. Por exemplo, alinhadores que analisam nomes de conceitos são mais adequados para ontologias que possuem termos parecidos para descrição de suas entidades. O mesmo alinhador pode não gerar bons resultados ao alinhar duas versões de uma mesma ontologia em diferentes idiomas.

Para tratar o problema de calibragem de alinhadores, este é modelado como um problema de otimização e, assim, é realizada uma discretização do espaço de soluções para reduzir o espaço de busca. Com isso, este trabalho

[6] http://wordnet.princeton.edu/
[7] http://www.dbpedia.org

apresenta uma abordagem heurística para otimizar o alinhamento gerado à partir dos alinhadores escolhidos para alinhar duas ontologias.

Dentre as metaheurísticas existentes, optou-se por uma abordagem heurística uni-objetivo baseada em populações utilizada em conjunto com uma heurística baseada em soluções únicas. Metaheurísticas baseadas em populações exploram o espaço de soluções através de uma diversificação das soluções iniciais. Para problemas em que podem existir muitos ótimos locais, a exploração do espaço de soluções através de um processo de diversificação pode gerar melhores resultados do que através de um processo de intensificação da solução (como é o caso das metaheurísticas baseadas em soluções únicas). Para este trabalho, contudo, mostra-se que a introdução de um processo de intensificação na heurística baseada em populações, utilizando uma heurística baseada em soluções únicas, auxilia na convergência da solução. Uma abordagem uni-objetivo é aquela que possui uma única função objetivo, isto é, um único critério a ser maximizado ou minimizado.

Uma vez que este trabalho objetiva uma solução aplicável em tempo de execução, abordagens heurísticas baseadas em população possuem outras vantagens aplicacáveis a este problema. Primeiramente, são abordagens naturalmente paralelizáveis (Whitley, 2001), uma vez que os indivíduos podem ser avaliados de forma independente em cada nó paralelo. Além disso, seu comportamento evolutivo permite que a execução do algoritmo seja interrompida a qualquer momento para retornar a melhor solução encontrada até aquele instante. Esta pode ser uma propriedade desejável em cenários de alinhamento em tempo real (Bock et al, 2012).

Para avaliar esta proposta em comparação a outras propostas da literatura, optou-se por utilizar um *benchmark* para alinhamentos de ontologias amplamente utilizado por pesquisadores desse campo. Ao comparar os resultados em relação a uma mesma base de testes, é possível verificar com maior precisão a contribuição desta proposta e suas diferenças em relação às demais propostas da literatura. Para possibilitar essa comparação, uma ferramenta é desenvolvida, a qual permite a utilização dessa proposta em trabalhos futuros. A comparação entre as propostas é realizada com base em medidas sugeridas pelo *benchmark*. Ainda, a análise das propostas é realizada em relação a características propostas na literatura.

1.4 Organização da tese

Esta tese está dividida em sete capítulos, sendo o capítulo 1 esta introdução. O capítulo 2 apresenta o problema de calibragem de alinhadores, as definições que são utilizadas neste trabalho e apresenta o estado da arte do meta-alinhamento de ontologias. A abordagem heurística proposta neste estudo é apresentada no capíutlo 3. Neste capítulo também é apresentada a demonstração do comportamento da abordagem em diferentes configurações com o objetivo de confirmar que a abordagem é capaz de encontrar boas soluções nestas diferentes configurações.

Este trabalho apresenta a ferramenta GNoSIS+, a qual pode ser utilizada para alinhar ontologias com base na abordagem heurística proposta. Os módulos da ferramenta são apresentados no capítulo 4.

A ferramenta GNoSIS+ é utilizada para alinhar vários exemplos de ontologias disponíveis em um *benchmark* utilizado por pesquisadores do campo de alinhamento de ontologias. Os resultados alcançados são apresentados no capítulo 5 e, no capítulo 6, são comparados com outras ferramentas que utilizam abordagens similares ao GNoSIS+.

Por fim, o capítulo 7 apresenta as conclusões do trabalho, descrevendo as contribuições dessa pesquisa e suas limitações. Ainda, a pesquisa apresentada neste trabalho mostra que há espaço para melhorias, as quais são apresentadas como trabalhos futuros também no capítulo 7.