Uma Estratégia para o Alinhamento Taxonômico de Ontologias

Carolina Howard Felicíssimo, Karin Koogan Breitman
Departamento de Informática – PUC-RIO
Rua Marquês de São Vicente, 225, CEP: 22453-900, Rio de Janeiro – RJ – Brasil
e-mail: [cfelicissimo, karin]@inf.puc-rio.br

Resumo

Com a evolução da Web para a Web Semântica, onde informações estarão disponibilizadas de forma que máquinas possam processá-las, surge a necessidade de mecanismos que garantam a interoperabilidade entre aplicações. Uma das possibilidades, que já vem sendo empregada na Web, é o uso de agentes de software. Estes programas atuam em ambientes abertos e heterogêneos e, portanto, precisam de informações âncoras para que possam efetivamente colaborar. A proposta mais comum na literatura é o uso de ontologias. Uma vez que diferentes ontologias estejam disponíveis na Web, o problema passa a ser como compatibilizar ontologias de modo a garantir a comunicação entre agentes de software. Na literatura este problema é tratado como alinhamento de ontologias. Neste artigo apresentamos uma estratégia automática para o alinhamento taxonômico de ontologias.

1. Introdução

Atualmente na Web tem-se um grande volume de informações disponibilizadas sem uma forma adequada para a representação de conhecimento. Desta maneira, o conteúdo das páginas Web é passível de ser processado apenas por humanos; máquinas não obtêm suporte explícito para este tipo de tarefa. Frente esta dificuldade, pesquisadores da indústria e da academia vêm explorando a possibilidade de criar uma Web Semântica. Nesta nova Web, informações estarão organizadas de forma que máquinas processem e integrem seus recursos de maneira inteligente, possibilitando buscas mais rápidas e precisas, e facilitando a comunicação entre seus dispositivos heterogêneos [1].

Ontologias, definidas como "*especificações explícitas e formais de conceitos compartilhados*" [2], vêm sendo utilizadas para fornecer suporte ao processamento das informações disponíveis na Web por máquinas. Através de ontologias, informações são estruturadas utilizando-se um vocabulário livre de ambigüidades e com um formalismo passível de processamento automático.

A Web do futuro será composta de várias ontologias pequenas e altamente contextualizadas, desenvolvidas localmente por engenheiros de software e não por especialistas em ontologias [3]. Neste cenário, além do processamento de informações por máquinas, é desejável que estas consigam suporte para a interoperabilidade de ontologias.

O objetivo desse artigo é atacar o problema de compatibilidade de ontologias para que os agentes de software que atuarão na Web Semântica possam efetivar colaborações. Nesse trabalho, procura-se seguir algumas das estratégias aplicadas à representação *OWL* [4], linguagem atual padrão da W3C, adotadas na detecção de similaridades. Uma estratégia para o alinhamento taxonômico de pares de ontologias no contexto da Web Semântica é apresentada. Os mecanismos de interoperabilidade de ontologias são apresentados na Seção 2. Na Seção 3, alguns trabalhos relacionados à Seção 2 são mencionados. Na Seção 4, a estratégia para o alinhamento taxonômico automático de ontologias é descrita. Na seção 5, um estudo de caso é relatado. Por fim, na Seção 6, as conclusões são apresentadas.

2. Interoperabilidade de Ontologias

Atualmente, existem algumas estratégias para compatibilidade de ontologias para a Web semântica, entre elas: (1) combinação [5] (2) alinhamento [5], (3) mapeamento [6], (4) integração [7], entre outras.
Na *combinação* de ontologias tem-se como resultado a versão das ontologias originais combinadas em uma ontologia única com todos seus termos juntos, sem a definição clara de suas origens. Normalmente as ontologias originais descrevem domínios similares ou de sobreposição.
No *alinhamento* de ontologias tem-se como resultado as duas ontologias originais separadas, mas com as ligações estabelecidas entre elas, permitindo que as ontologias alinhadas reusem as informações uma das outras. O alinhamento normalmente é realizado quando as ontologias são de domínios complementares.
No *mapeamento* de ontologias tem-se como resultado uma estrutura formal que contém expressões que fazem a ligação de conceitos de um modelo em conceitos de um segundo modelo. Este mapeamento pode ser usado para transferir instâncias de dados, esquemas de integração, esquemas de combinação e tarefas similares.
Na *integração* de ontologias tem-se como resultado uma ontologia única criada pela montagem, extensão, especialização ou adaptação de outras ontologias de assuntos diferentes. Na integração de ontologias é possível identificar as regiões que foram criadas a partir das ontologias originais.
Nesse trabalho, busca-se a identificação de termos equivalentes em aplicações complementares de forma a permitir a negociação de suas informações. Desta maneira, o mecanismo de alinhamento de ontologias é o utilizado.

3. Trabalhos Relacionados

A interoperabilidade de ontologias vem sendo estudada por diferentes pesquisadores. Naturalmente, algumas abordagens distintas têm sido exploradas. Por exemplo, o sistema GLUE [8] faz uso de estratégias de aprendizagem múltiplas para encontrar os *mapeamentos* semânticos entre duas ontologias. O serviço semi-automático OntoMerge [9] realiza a *combinação* de ontologias pela união de seus axiomas. O serviço Articulation Service [10] realiza o *mapeamento* de duas ontologias.
Para o trabalho com múltiplas e extensas ontologias, é proposto o conjunto de ferramentas PROMPT [6]. Dentre as ferramentas deste conjunto, tem-se: iPROMPT, uma ferramenta interativa para *combinação* de ontologias; AnchorPROMPT, uma ferramenta automática baseada em grafos para *alinhamento* de ontologias; PROMPTFactor, uma ferramenta para *extração* de partes de ontologias e PROMPTDiff, uma ferramenta para *identificação de diferenças* entre duas versões da mesma ontologia.
Na comunidade de Banco de Dados, o problema de *mapeamento* dos diferentes esquemas de bancos é antigo. Entre as soluções possíveis estão o uso de conversores, mediadores e técnicas de mapeamento. Além disso, soluções específicas para ontologias já estão sendo estudadas por essa comunidade [11].
Apesar dos trabalhos realizados para garantir mecanismos que suportem a interoperabilidade de ontologias, ainda não foi encontrada uma solução razoável para o alinhamento de ontologias, prioridade desse trabalho. Na próxima seção é apresentada uma estratégia para alinhamento de ontologias baseada em técnicas utilizadas pela Engenharia de Software. A estratégia proposta é validada através de um protótipo, o Componente para Alinhamento Taxonômico de Ontologias - CATO [12]. O CATO alinha automaticamente, ou seja, sem a intervenção de usuários, as taxonomias das ontologias de entrada. O CATO

foi totalmente implementado em Java [13] e usa a *API* Jena [14] específica para o tratamento de ontologias. A versão on-line do CATO está disponível para uso público em: http://cato.les.inf.puc-rio.br.

4. A Estratégia

Alinhar os termos de diferentes ontologias continua um problema em aberto e que precisa ser resolvido para viabilizar uma série de promessas da Web semântica. Uma das mais crítica é a necessidade de garantir a comunicação automática entre agentes de software em aplicações semânticas permitindo a cooperação, i.e., compartilhamento e reutilização, da informação disponibilizada.

Na tentativa de contribuir para a evolução da Web, propomos uma estratégia para o alinhamento taxonômico de ontologias. Devido ao fato do componente central em uma ontologia ser sua taxonomia [15], no primeiro momento, apenas os conceitos com relacionamentos de especialização entre eles, i.e., relacionamentos do tipo "é-um", entre duas ontologias de entrada são investigados. Na estratégia proposta, ilustrada na Figura 1, o alinhamento é obtido ao final de três etapas executadas seqüencialmente. A estratégia tem como entradas duas ontologias e como saída uma ontologia única, representando as ontologias originais alinhadas.

A primeira etapa da estratégia faz uso de comparação lexical entre os conceitos das ontologias de entrada e mecanismo de poda estrutural dos conceitos associados como condição de parada. O objetivo desta etapa é realizar a comparação lexical dos conceitos das ontologias de forma "mais inteligente", com o enriquecimento de informações, a procura de conceitos iguais lexicalmente e com o mesmo significado, i.e., conceitos equivalentes. A identificação automática da semântica dos conceitos comparados é conseguida de forma a satisfazer a condição de generalização, nomes iguais dos conceitos associados encontrados dois níveis hierárquicos acima do conceito comparado, e a condição de especialização, nomes iguais das instâncias cadastradas nas ontologias. Satisfeitas estas duas condições, os conceitos equivalente identificados são alinhados já nesta etapa. Os resultados da etapa são as ontologias de entrada enriquecidas com os possíveis alinhamentos. Estas ontologias enriquecidas são transformadas em arquivos do tipo XML [16] onde apenas suas hierarquias são representadas.

A segunda etapa da estratégia compara estruturalmente as hierarquias das ontologias, identificando as similaridades entre suas sub-árvores comuns. Para comparação de árvores, existem algoritmos de busca de similaridades estruturais, utilizados em várias aplicações de Engenharia de Software, tais como: TreeDiff [17], TreeToTree [18], TreeMatcher [19], entre outros. O algoritmo do TreeDiff descrito em [20] foi o escolhido por satisfazer os requisitos para a busca de similaridades estruturais e ter sua implementação disponibilizada. A comparação estrutural do TreeDiff utiliza grupos de equivalência, identificados tanto pela comparação lexical quanto pela comparação estrutural. De início, os conceitos comparados com o mesmo nome são identificados pela comparação lexical e, em seguida, as informações da quantidade de filhos (sub-conceitos) e os conceitos equivalentes que estes filhos possuem são analisadas pela comparação estrutural. O resultado desta etapa da estratégia são os conceitos dos grupos de equivalência identificados como equivalentes.

A terceira etapa refina os resultados da etapa anterior classificando aqueles conceitos identificados como equivalentes em *bem equivalentes* ou *pouco equivalentes*, de acordo com um percentual de similaridade pré-fixado. O resultado desta etapa são os conceitos classificados como bem equivalentes. Tais conceitos são adicionados nas ontologias

resultantes da primeira etapa da estratégia e estas unidas em uma ontologia única, resultado final da estratégia.

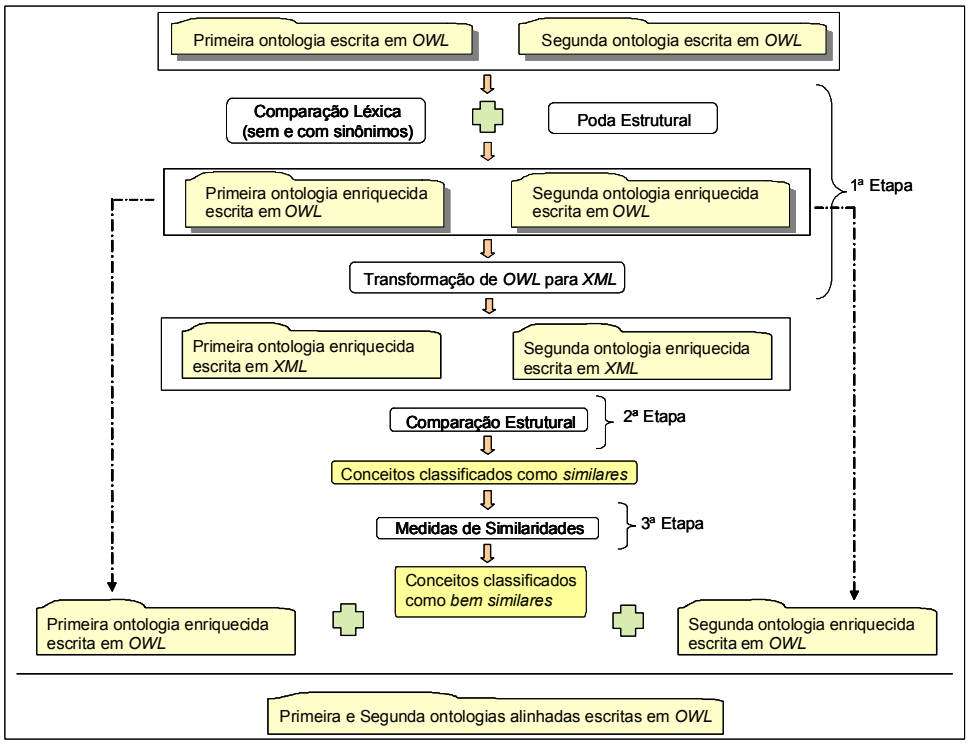

Figura 1. Estratégia para o Alinhamento Taxonômico de Ontologias.

O fato do resultado final da estratégia ser uma ontologia única foi uma decisão de implementação. As ontologias originais continuam sendo reconhecidas pela identificação de seus *namespaces* e existe a ligação entre os conceitos equivalentes na ontologia única permitindo, assim, a reutilização e o compartilhamento das informações comuns.

5. Estudo de Caso

Duas ontologias independentes, criadas por diferentes grupos e disponíveis publicamente na Web, foram escolhidas como exemplo para este estudo de caso. A primeira ontologia é a *CMU RI Publications* [21] do projeto de pesquisa *Agent Transaction Language for Advertising Services* (ATLAS) [22] da Universidade Carnegie Mellon [23]. A segunda ontologia é a *General University Ontology* [24] de uma das empresas participantes do grupo de trabalho da Web-Ontology da W3C [25]. As ontologias escolhidas possuem diferenças estruturais e no número de seus conceitos.

A Figura 2 ilustra as hierarquias das duas ontologias comparadas, com as abreviações O1 representando a primeira ontologia (*CMU RI Publications)* e O2 representando a segunda ontologia (*General University Ontology*). A primeira ontologia possui vinte e cinco conceitos no total, e a segunda, duzentos e vinte e cinco.

Após a análise manual das ontologias comparadas, identificamos que apenas oito conceitos de cada uma, apontados pelas setas na Figura 3, poderiam ser alinhados. As setas finas da figura apontam para os conceitos que poderiam ser alinhados se fosse permitida a intervenção humana e as setas mais grossas apontam para os conceitos que são alinhados, de forma totalmente automática, pelo CATO.

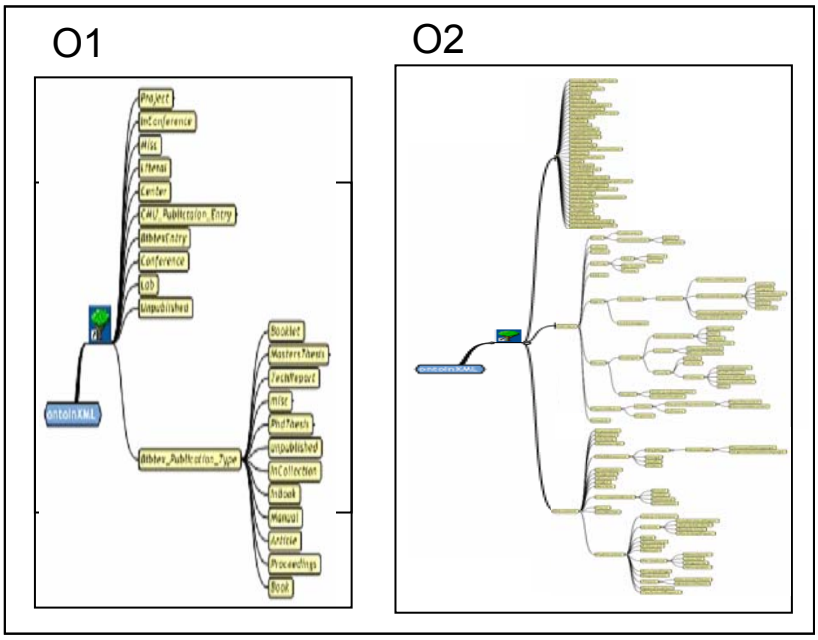

Figura 2. Hierarquias das ontologias comparadas.

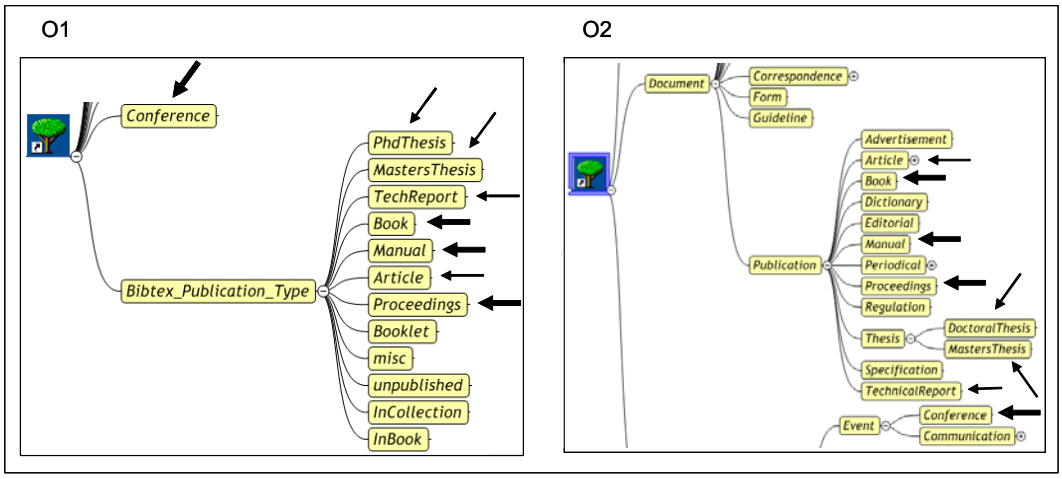

Figura 3. Conceitos que poderiam ser alinhados nas ontologias comparadas.

5.1. Primeira Etapa: Comparação Lexical com Uso de Sinônimos e Mecanismo de Poda Estrutural como Condição de Parada

A etapa inicial da estratégia compara lexicalmente os conceitos das ontologias, fazendo uso de sinônimos durante esta comparação. Um banco de sinônimos com acesso automático disponibilizado foi criado para uso do CATO. Neste banco, novos sinônimos são adicionados, manualmente e sistematicamente, ao longo de seu uso, de forma a diluir o custo de sua construção. Para esse estudo de caso, os seguintes sinônimos foram cadastrados no banco: "TechReport" como sinônimo de "TechnicalReport", "TechnicalReport como sinônimo de "TechReport", "PhdThesis" como sinônimo de "DoctoralThesis" e "DoctoralThesis" como sinônimo de "PhdThesis".

Os conceitos representados com os mesmos nomes nas ontologias comparadas ("Conference", "MastersThesis", "Book", "Manual", "Article" e "Proceedings") e os com seus sinônimos cadastrados são comparados, mas como não satisfazem as condições de poda, porque possuem diferentes conceitos nos dois níveis hierárquicos superiores, não são alinhados.

5.2. Segunda Etapa: Comparação Estrutural Usando uma Implementação do Algoritmo *TreeDiff*

Esta etapa da estratégia compara as estruturas de ontologias. A Figura 4 ilustra as hierarquias das ontologias comparadas com seus grupos de equivalências circulados.

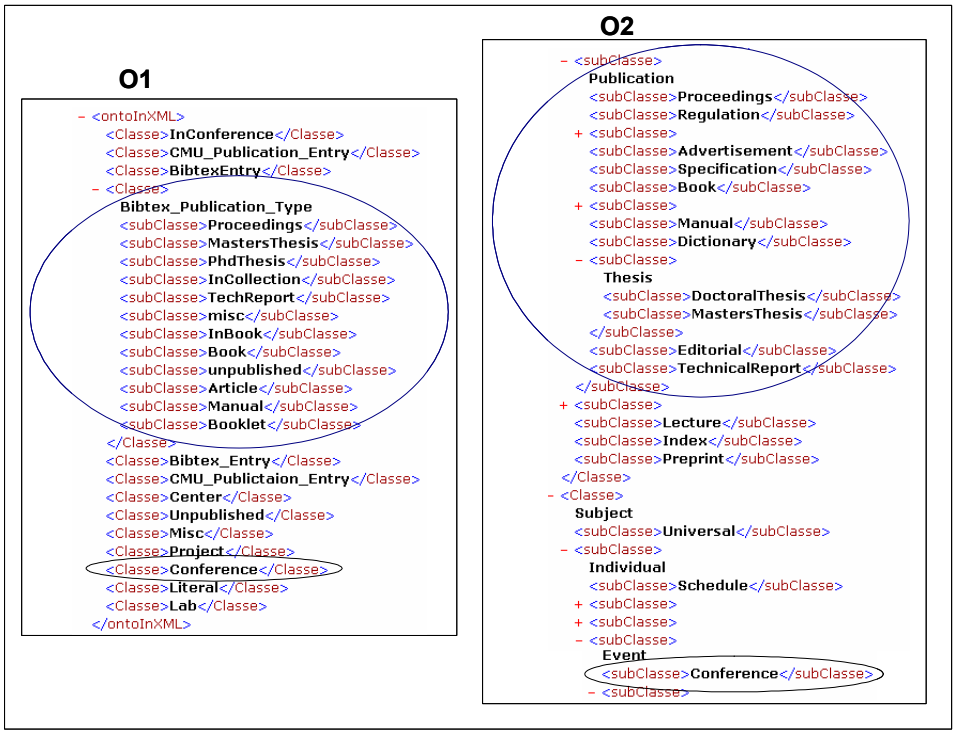

Figura 4. Parte das hierarquias das ontologias comparadas.

Os primeiros grupos de equivalência, representados pelos círculos superiores na Figura 4, são formados pela igualdade lexical dos conceitos nomeados "Proceedings" de ambas ontologias e pela similaridade estrutural entre seus super-conceitos (um nível hierárquico acima). Após esses grupos de equivalências serem formados, todos os conceitos dentro deles são comparados a procura de novas igualdades lexicais e similaridades estruturais. Assim, os conceitos nomeados "Book" e "Manual" são identificados como equivalentes.
Devido à similaridade estrutural e igualdade de seus nomes, os conceitos "Conference" de ambas ontologias são também identificados e os novos grupos de equivalência, representados pelos círculos inferiores na Figura 4, são formados.
No entanto, outros conceitos com nomes iguais, tais como os conceitos "MastersThesis" e "Article", presentes em ambas ontologias, não são identificados como equivalentes porque suas estruturas hierárquicas possuem diferenças.
No final desta etapa da estratégia, os conceitos "Proceedings", "Book", "Manual" e "Conference" das ontologias comparadas são identificados como equivalentes. Oito novos conceitos poderiam ser também identificados como equivalentes se fosse permitida a intervenção do usuário (os conceitos "Article" e "MastersThesis" de ambas ontologias, "PhdThesis" de O1 com "DoctoralThesis" de O2 e "TechReport" de O1 com "TechnicalReport" de O2).

5.3. Terceira Etapa: Uso de Medidas de Similaridades para os Ajustes Finos

A etapa final da estratégia classifica os conceitos equivalentes, identificados na etapa anterior, como *bem equivalentes* ou *pouco equivalentes* de acordo com um percentual pré-

definido para medida de similaridade entre ontologias. Atualmente estamos trabalhando com o percentual de similaridade igual a setenta e cinco por cento. Este resultado foi obtido empiricamente, através da análise e refinamento de resultados obtidos em experimentos prévios de alinhamento de ontologias.
O CATO só alinhará os conceitos com percentuais de similaridades maior ou igual a setenta e cinco por cento, ou seja, os conceitos classificados como bem equivalentes. A Figura 5 ilustra os resultados dos percentuais de similaridades calculados pelo CATO. No final desta etapa da estratégia, os conceitos bem equivalentes "Proceedings", "Book", "Manual" e "Conference", de ambas ontologias, são os conceitos alinhados pelo CATO.

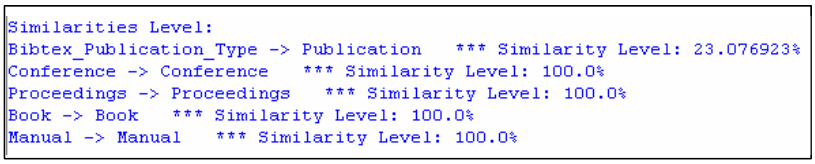

Figura 5. Percentuais de similaridades calculados pelo CATO.

6. Conclusão

Nesse trabalho abordamos parte do problema de alinhamento de ontologias. Apresentamos uma estratégia para o alinhamento taxonômico de ontologias baseada na comparação lexical e estrutural, e uso de medidas de similaridades para tomada de decisão. Tais soluções, largamente utilizadas na Ciência da Computação, foram customizadas para o tratamento com ontologias.
Hoje, a Internet possui mais de quatro bilhões de páginas [26]. Neste ambiente, o alinhamento manual ou até mesmo o semi-automático passa a ser inviável. O CATO, resultado da estratégia apresentada, está disponível na Internet para uso público, respondendo como uma ferramenta de alinhamento automático de taxonomias de ontologias. Não foi encontrada outra ferramenta que realize também tal alinhamento.
O CATO traz bons resultados quando aplicado com ontologias de domínios complementares, similares e de sobreposição. Nestas ontologias, os conceitos equivalentes estão normalmente próximos estruturalmente e são identificados com os mesmos nomes ou com seus sinônimos. Desta maneira, a identificação dos grupos de equivalência é mais precisa e, conseqüentemente, o alinhamento também.
Como trabalhos futuros, planeja-se a possibilidade de alinhamento de mais termos das ontologias comparadas além da adição de novos recursos que melhorem tanto a comparação lexical quanto a estrutural. Mais estudos de caso devem ser realizados.

Referências

1. Berners-Lee, T., Lassila, O. Hendler, J.: *The Semantic Web*, disponível em: http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html, acesso em Maio de 2004.
2. Gruber, T. R.: *A translation approach to portable ontology specifications,* Knowledge Acquisition, Vol.5, pp. 199-220 - 1993.
3. Hendler, J.: *Agents and the Semantic Web*, IEEE Intelligent Systems, Março/Abril - pp.30-37 - 2001
4. Dean, M. et. al. *OWL Web Ontology Language Reference.* Disponível em: <http://www.w3.org/TR/owl-ref/>. Acesso em Maio de 2004.
5. Noy, N. F., Musen, M. A.: *SMART: Automated Support for Ontology Merging and Alignment*, Banff Workshop on Knowledge Acquisition, Modeling, and Management, Banff, Alberta, Canada. 1999.

6. Noy, N. F., Musen, M. A.: *The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping*, International Journal of Human-Computer Studies, 2003.
7. Pinto, S. H., Gómez-Pérez, A., Martins, J. P.: *Some Issues on Ontology Integration.,* In Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends, 1999.
8. Bouquet, P., Serafini, L., Zanobini, S.: *Semantic Coordination: A New Approach and an Application*. *In Proceedings of the 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Sundial Resort, Sanibel Island, Florida, USA.*
9. *OntoMerge - Ontology Translation by Merging Ontologies.* Disponível em: <http://onto.cs.yale.edu:4040/ontoMerge.html>. Acesso em: Junho de 2004.
10. *Articulation Service.* Disponível em: <http://codip.grci.com/Tools/ArtiServicePage.html>. Acesso em: Junho de 2004.
11. Moreira, M. M. Integração Semântica de Sistemas de Informação. Dissertação de Mestrado, PUC-Rio, 2003.
12. Felicíssimo, C. H. CATO – Componente para Alinhamento Taxonômico de Ontologias. Projeto de Programação do Departamento de Informática da PUC-Rio para o curso de Mestrado. Dezembro de 2003. Disponível em: <http://cato.les.inf.puc-rio.br>. Acesso em: Setembro de 2004.
13. *The Java Language Specification. Second Edition*. Gosling, J., et. al. 2000. Disponível em: <http://java.sun.com/docs/books/jls/second_edition/html/j.title.doc.html>. Acesso em Junho de 2004.
14. *Jena 2 Ontology API*. Disponível em: <http://jena.sourceforge.net/ontology/>. Acesso em Junho de 2004.
15. Doan, A., Dhamankar, R., Domingos, P., Halevy, A.: *Learning to match ontologies on the Semantic Web*, *The VLDB Journal — The International Journal on Very Large Data Bases, Volume 12, Issue 4, November, 2003. Pages: 303 – 319. ISSN:1066-8888.*
16. *Extensible Markup Language (XML)*. Disponível em: <http://www.w3.org/XML/>. Acesso em Junho de 2004.
17. Wang, J.: *An Algorithm for Finding the Largest Approximately Common Substructures of Two Trees*, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, Nº. 8, pp. 889-895, Aug. 1998.*
18. TAI, K.,C. *The tree-to-tree correction problem. Journal of the ACM, 26(3), P. 422-433, 1979*.
19. *TreeMatcher toolkit for comparing two ordered or unordered trees*. Disponível em: <http://www.cis.njit.edu/~discdb/treematcher.html>. Acesso em: Junho de 2004.
20. Bergmann, U.: Evolução de Cenários Através de um Mecanismo de Rastreamento Baseado em Transformações, Tese de Doutorado, PUC-Rio, 2002.
21. *CMU RI Publications.* Disponível em <http://www.daml.ri.cmu.edu/ont/homework/cmu-ri-publications-ont.daml>. Acesso em: Junho de 2004.
22. *Agent Transaction Language for Advertising Services* (ATLAS). Disponível em: <http://www.daml.ri.cmu.edu>. Acesso em: Junho de 2004.
23. *Universidade Carnegie Mellon.* Disponível em: <http://www.cmu.edu/>. Acesso em: Junho de 2004.
24. *General University Ontology*. Disponível em: <http://www.mondeca.com/owl/moses/univ.owl>. Acesso em: Junho de 2004.
25. Mondeca S.A. *A Semantic Knowledge company*. Disponível em: <http://www.mondeca.com>. Acesso em: Junho de 2004.
26. Google. Disponível em: <http://www.google.com/>. Acesso em Setembro de 2004.