Uma proposta para análise de similaridade entre documentos XML e ontologias em OWL

Rodrigo Perozzo Noll, Deise de Brum Saccol, Nina Edelweiss

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{rodrigo.noll, deisesaccol}@gmail.com, nina@inf.ufrgs.br

Abstract. *Document and schema matching is relevant in many scenarios, such as data integration, data warehousing, and semantic query processing. Existent approaches are based on lexical or semantic similarity analysis between the representations. However, there are few approaches for ontology and document matching, and the existent ones usually demand the intervention of a specialist. In this context, this paper presents an approach for similarity analysis between XML files and ontologies. The paper also presents a tool that verifies its viability and a case study that shows the efficacy and the good results provided.*

Resumo. *O casamento de documentos e esquemas é aplicável em diversos cenários, como integração de dados, data warehouse, e processamento semântico de consultas. A maioria das propostas baseia-se na similaridade léxica ou semântica entre as representações. No entanto, existem poucas propostas para casamento de documentos XML e ontologias e, na maioria das vezes, exigem intervenção de um especialista. Neste contexto, este trabalho apresenta uma proposta automática para análise de similaridade entre arquivos XML e ontologias. O artigo também descreve uma ferramenta que comprova a viabilidade e um estudo de caso que demonstra a eficácia e os bons resultados obtidos.*

1. Introdução

O casamento de documentos é uma operação que avalia a similaridade entre esquemas através do mapeamento de seus conteúdos. No entanto, combinar dados de domínios distintos pode acarretar problemas, como, por exemplo, diferentes formatos de representação e inconsistências semânticas [Ding 2002]. Desta forma, para avaliar corretamente o grau de similaridade entre documentos, é interessante considerar tanto perspectivas léxicas quanto semânticas [Maedche 2002]. A perspectiva léxica avalia as relações entre os termos, comparando suas cadeias de caracteres. Já a perspectiva semântica concentra-se no significado e na correlação conceitual entre estes termos. Com o objetivo de identificar o relacionamento semântico entre documentos, é necessário um formalismo lógico. Neste contexto, ontologias são apropriadas para a modelagem conceitual [Gruber 1993].

Existem algumas abordagens para análise de similaridade de documentos (baseadas na estrutura [Nierman 2002] e no conteúdo [Baeza-Yates 1999]), de esquemas [Rahm 2004], e de esquemas com ontologias [Aumueller 2005]. No entanto, estas abordagens são geralmente baseadas na similaridade léxica entre os elementos. Além disso, o casamento de documentos com ontologias ainda é pouco explorado. O objetivo da proposta apresentada neste artigo é a avaliação do grau de similaridade de todos os elementos definidos nas duas representações (XML e ontologia), considerando tanto a perspectiva léxica quanto a semântica. Para tanto, foi desenvolvida uma ferramenta que automatiza esta avaliação, possibilitando definir qual ontologia melhor descreve um documento XML.

Este artigo está estruturado da seguinte forma: a seção 2 apresenta a proposta de avaliação de similaridade entre documentos XML e ontologias. A Seção 3 apresenta a implementação desenvolvida para simular a proposta apresentada. A Seção 4 apresenta os resultados obtidos através de um estudo de caso. As conclusões e trabalhos futuros são apresentados na Seção 5.

2. Proposta de avaliação de similaridade entre documentos XML e ontologias

A avaliação da similaridade vem sendo alvo de constante pesquisa, na qual a literatura sugere três passos como principais [Madhavan 2001]:

1. *Normalização*: elementos semanticamente equivalentes podem ter nomes diferentes em esquemas distintos. Neste passo, sugere-se utilizar um Tesauro com termos de uma linguagem comum ou referências de domínio específico.
2. *Categorização***:** o objetivo deste passo é separar elementos em classes, visando reduzir o número de comparações entre elementos distintos.
3. *Comparação***:** consiste na definição de um coeficiente de similaridade, computado entre os elementos em suas respectivas categorias.

Este trabalho considera tanto a perspectiva léxica quanto a semântica.

Análise de similaridade léxica. A literatura sugere duas principais abordagens:

- *Edit Distance* [Levenshtein 1966]: avalia duas seqüências de caracteres pelo número mínimo de operações necessárias para transformar uma cadeia em outra;
- *Stemmer* [Stemmer 2007]: avaliação de seqüência de caractere pela redução de uma palavra ao seu radical.

Em [Kantrowitz 2000] é apresentado um comparativo entre algoritmos de análise léxica, incluindo *Edit Distance* e *Stemmer*. O algoritmo *Edit Distance* possui um melhor desempenho em situações onde não existe uma avaliação da grafia correta dos elementos avaliados, como, por exemplo, em textos que contém erros de digitação. Como a organização taxonômica dos elementos é definida *a priori*, evitando-se erros gramaticais, sugere-se a utilização do algoritmo de *Stemmer* para avaliação de similaridade léxica.

Análise de similaridade semântica. A segunda perspectiva corresponde à avaliação semântica entre os termos. Durante o passo de normalização, sugere-se a utilização de um Tesauro para avaliar relações terminológicas entre conceitos. A *WordNet* [WordNet 2007] representa a maior base de dados léxica da língua inglesa e relaciona substantivos, verbos, adjetivos e advérbios. Estes elementos estão agrupados em um conjunto de sinônimos cognitivos e relacionados sobre a perspectiva léxica e semântica.

Outro procedimento existente é o *Taxonomic Overlap* [Maedche 2002], que corresponde à comparação taxonômica entre os elementos avaliados. Esta comparação não avalia o elemento individualmente, mas sim o contexto em que este se encontra com relação aos demais. Para a definição do grau de similaridade, aplica-se uma medida conhecida como *Coeficiente de Jaccard* (CJ) [Manning 1999], onde a similaridade de dois conjuntos é dada pela divisão da cardinalidade resultante das operações de intersecção e união.

No exemplo da Figura 1, suponha que o nodo A seja lexicamente equivalente ao nodo 1 e o nodo B seja lexicamente equivalente ao nodo 3. Para avaliar o *Taxonomic Overlap* estabelecido pela hierarquia A, B e C, sabemos que a união é três (elementos A, B e C) e que a intersecção é dois (elemento A equivalente a 1 e B equivalente a 3). O grau de similaridade estabelecido pelo CJ será dois (intersecção) dividido por três (união), ou seja, o valor 0,667 (coeficiente entre os valores 0 - má combinação, e 1 - combinação perfeita).

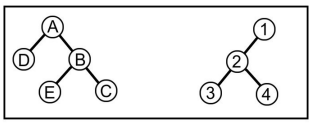


Figura 1 – Cenário de elementos

3. *The Matcher*: Uma ferramenta para avaliação de similaridade

A ferramenta *The Matcher*[1] comprova a viabilidade da proposta de avaliação da similaridade léxica e semântica entre documentos XML e ontologias. O objetivo é verificar, dentre um conjunto de ontologias, qual a que melhor descreve um documento XML.

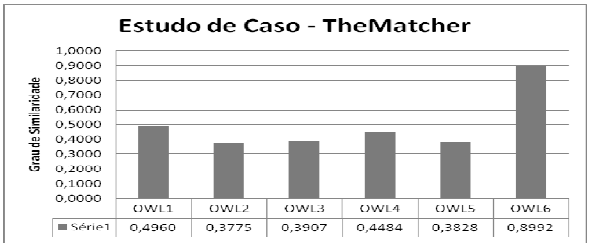
A primeira etapa corresponde à normalização e categorização. Para cada documento (XML e OWL), foi realizado um mapeamento dos elementos que os compõem. Percorre-se o documento e armazena-se, para cada elemento, seu radical (chave) e uma lista com o nome completo do elemento. A próxima etapa compreende a expansão da categorização de seus sinônimos, isto é, adicionar à lista todas as palavras sinônimas, recuperadas através da WordNet. Após a expansão, a ferramenta verifica a correspondência léxica (utilizando *Stemmer)* entre os elementos da lista de cada documento.

Por fim, ocorre a avaliação utilizando o *Taxonomic Overlap*, considerando a organização estabelecida desde a raiz do documento XML a cada uma de suas folhas. Para cada elemento que compõe um conjunto raiz-folha, sabe-se da existência ou não de correspondente na ontologia (definido durante a normalização). Para obter o grau de similaridade (CJ), tem-se como união o número total de elementos do conjunto raiz-folha do XML e como intersecção o número de relações em que existe correspondência.

4. Avaliação da Ferramenta

Para avaliar os resultados obtidos, optou-se por realizar um estudo de caso com o objetivo de avaliar o grau de similaridade de distintos documentos utilizando a ferramenta *The Matcher*. A questão a ser respondida é se o valor de similaridade de documentos de mesmo domínio é maior do que documentos de domínios distintos. A métrica utilizada para responder a questão é o resultado obtido pela utilização da ferramenta sobre distintos modelos.

As variáveis independentes (entrada do estudo de caso) são os documentos XML e OWL. Como variável dependente (saída), definiu-se o grau de similaridade entre dois documentos. A instrumentação foi obtida a partir de duas fontes: [Protegé Library 2007] para ontologias e [XML Résumé Library 2007] para documentos XML que descrevem currículos. As ontologias utilizadas definem domínios distintos, dentre eles: **OWL1** (descrição de pesquisa acadêmica); **OWL2** (propriedades de aminoácidos); **OWL3** (vinhos); **OWL4** (relacionamento entre animais de estimação e pessoas); **OWL5** (turismo); **OWL6** (currículo e carreira profissional). Os resultados são apresentados na Figura 2.



[1] A ferramenta *The Matcher* está disponível para *download,* juntamente com seu código fonte, pela URL http://www.inf.ufrgs.br/~deise/poster2007-Rodrigo/TheMatcher.zip.

Fig. 2. Resultados obtidos pela ferramenta

Avaliando o resultado obtido entre o documento XML (currículo) e a ontologia OWL6 (currículo e carreira profissional), mesmo de fontes totalmente diferentes, o grau de similaridade obtido é de 89,92%, isto é, quase 90% das estruturas definidas em ambos os documentos são lexicamente e semanticamente similares. Este fenômeno não foi observado nas demais ontologias (domínios distintos), todas com graus de similaridade inferiores a 50%. Com base na avaliação dos dados obtidos, consegue-se verificar que o grau de similaridade obtido pela proposta para documentos de mesmo domínio é consideravelmente superior que o grau obtido a partir de documentos de domínios diferentes.

5. Conclusões e trabalhos futuros

Este trabalho apresentou uma proposta para análise de similaridade entre documentos XML e ontologias OWL, adotando boas práticas definidas na literatura. Para avaliar a viabilidade e eficácia da proposta, desenvolveu-se uma ferramenta e um estudo de caso. Como trabalho futuro, pretende-se ampliar o conjunto e variar o domínio dos documentos e das ontologias testadas. Também se pretende incorporar a ferramenta ao *DetVX* [Saccol 2007], um ambiente para detecção e gerenciamento de réplicas e versões de documentos XML em cenários *peer-to-peer*. A proposta e a ferramenta *The Matcher* vão auxiliar na etapa de descoberta do domínio de conhecimento (ontologia) que descreve um conjunto de documentos XML.

6. Referências

[Aumueller 2005] Aumueller, D.; Hong-Hai, D. "Schema and ontology matching with COMA++". In: Proceedings of the 2005 ACM SIGMOD, 906 – 908, 2005.

[Baeza-Yates 1999] Baeza-Yates, R.A. e Ribeiro-Neto, B.A.. "Modern Information Retrieval". ACM Press / Addison-Wesley, 1999.

[Ding 2002] 2002] Ding, Y.; Foo, S. "A review of ontology mapping and evolving". Journal of Information Science, v. 28, n. 5,p. 375-388, October 2002.

[Gruber 93] Gruber, T.R. "Towards Principles for the Design of Ontologies Used for Knowledge Sharing". International Journal of Human and Computer Studies, 43/(5/6):907-928,1993.

[Kantrowitz 2000] Kantrowitz, M.; Mohit, B.; Mittal, V. "Stemming and its effects on TFIDF Ranking".

[Levenshtein 1966] Levenshtein, V. "Binary Codes capable of correcting deletions, insertions, and reversals". Cybernetics and Control Theory, 10(8):707–710, 1966.

[Madhavan 2001] Madhavan, J.; Bernstein, P. A.; Rahm, E. "Generic schema matching using Cupid". In: Proceedings of the 27th Very Large Data Bases, p. 48-58, 2001.

[Maedche 2002] Maedche, A.; Staab, S. "Measuring similarity between ontologies". In: Proceedings of the European Conference on EKAW, 2002.

[Manning 1999] Manning, C. D.; Schütze, H. "Foundations of Statistical Natural Language Processing". 1st ed. Cambridge, Massachusetts: MIT Press, 1999. 620 p.

[Nierman 2002] Nierman, A. e Jagadish, H.V. "Evaluating Structural Similarity in XML Documents". Proc. of the 5th Intl. Workshop on the Web and Databases, WebDB, 2002.

[Protegé Library 2007] Protegé Ontologies Library. Disponível em: http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary. Acesso em 27 de maio de 2007.

[Rahm 2004] Rahm E., Bernstein, P.A. "A survey of approaches to automatic schema matching". In: The VLDB Journal The International Journal on Very Large Data Bases, 2004.

[Saccol 2007] Saccol, D. B. ; Edelweiss, N. ; Galante, R.M. ; Zaniolo, C. "Managing XML Versions and Replicas in a P2P Context". In: Proc of the Nineteenth International Conference on Software Engineering and Knowledge Engineering, SEKE, 2007.

[Stemmer 2007] The Lancaster Stemming Algorithm. Disponível em: http://www.comp.lancs.ac.uk/computing/research/stemming. Acesso em 27 de maio de 2007.

[WordNet 2007] WordNet - A lexical database for the English language. Disponível em: http://wordnet.princeton.edu. Acesso em 27 de maio de 2007.

[XML Résumé 2007] XMLRésuméLibrary. Disponível em: http://xmlresume.sourceforge.net. Acesso em 27 de maio, 2007.