UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ESPECIALIZAÇÃO EM WEB E SISTEMAS DE INFORMAÇÃO

RODRIGO PEROZZO NOLL

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

Trabalho de Conclusão apresentado como requisito parcial para a obtenção do grau de Especialista

Profa. Dra. Renata de Matos Galante
Orientadora

Prof. Dr. Carlos Alberto Heuser
Coordenador do Curso

Porto Alegre, dezembro de 2007.

Noll, Rodrigo P.

Uma proposta para análise de similaridade entre documentos XML e ontologias definidas em OWL / Rodrigo Perozzo Noll – Porto Alegre: Curso de Especialização em WEB e Sistemas de Informação, 2007.

26 f.:il.

Trabalho de Conclusão de Curso (especialização) – Universidade Federal do Rio Grande do Sul. Curso de Especialização em WEB e Sistemas de Informação, Porto Alegre, BR – RS, 2007. Orientadora: Renata de Matos Galante.

1.Similaridade de Documentos. 2.Banco de dados. I. Galante, Renata de Matos. III. Título.

AGRADECIMENTOS

Aos meus pais, Carlos e Sônia, pela educação e exemplo de vida. Sua referência foi fundamental para estruturação de meu ser e meu caráter.

A minha esposa, Sandra, pela constante participação e carinho em cada detalhe de minha vida.

A minha orientadora, Professora Renata de Matos Galante, pelo incentivo, motivação e dedicação, fundamentais à realização deste trabalho.

A Deise de Brum Saccol, pelas importantes contribuições e sugestões necessárias para este trabalho.

Aos meus colegas do curso de pós-graduação, pela amizade e companheirismo.

A todos os professores e funcionários do Programa de Pós-Graduação em Computação da UFRGS.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CJ	Coeficiente de Jaccard
ED	Edit Distance
GQM	Goal-Question-Metric
OWL	Web Ontology Language
SC	Semantic Cotopy
UFRGS	Universidade Federal do Rio Grande do Sul
W3C	World Wide Web Consortium
XML	Extensible Markup Language

LISTA DE FIGURAS

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, o trabalho aqui apresentado objetiva propor uma estrutura que identifique o grau de similaridade entre estruturas definidas em XML e ontologias, considerando perspectivas léxicas e semânticas. Para tanto, foram avaliadas propostas encontradas na literatura para análise de similaridade entre esquemas, considerando ambas as perspectivas, visando identificar padrões e melhores práticas. Posteriormente, integraram-se estes padrões e boas práticas para avaliar a viabilidade prática da proposta. Com base no resultado da análise de viabilidade, desenvolveu-se um estudo de caso que avaliou os resultados obtidos. A principal contribuição deste trabalho diz respeito à automatização da avaliação de similaridade, possibilitando definir qual modelagem conceitual (ontologia) melhor descreve um documento XML.

Palavras-Chave: Similaridade, casamento de documentos, ontologia, XML.

A proposal to evaluate the similarity between XML documents and ontologies specified in OWL

ABSTRACT

Document and schema matching is relevant in many scenarios, such as data integration, data warehousing, and semantic query processing. However, there are a few approaches for ontology and XML document matching, and the existent ones usually demand the intervention of a specialist. In this context, this study presents an approach for similarity analysis between XML files and ontologies, considering lexical and semantic perspectives. For that end, some proposals for similarity evaluation between different schemas were evaluated to identify best practices and patterns. After that, these patterns and best practices were integrated to evaluate the current proposal's viability. Using the results acquired by the viability analysis, it was developed a survey. The main contribution of this work is the automation of the similarity evaluation that defines which conceptual model best describes a XML document.

Keywords: Similarity, document matching, ontology, XML.

1 INTRODUÇÃO

Atualmente, a manipulação do conhecimento é fundamental devido à crescente quantidade de informação gerada. O mesmo se reflete na quantidade e na velocidade em que o conhecimento é gerado. Este fato faz com que pessoas e organizações tenham que gerenciar seu conhecimento de modo mais eficaz. Combinar conhecimentos de domínios distintos pode acarretar problemas como, por exemplo, formatos de representação do conhecimento distintos e inconsistências semânticas, entre outros (DING, 2002).

O casamento de documentos é uma operação que avalia a similaridade entre esquemas através do mapeamento de seus conteúdos. A avaliação de similaridade é crítica em diversos cenários, como na identificação de pontos de integração em bancos de dados heterogêneos, no compartilhamento, integração e reuso de informação.

A integração ou determinação de equivalência das informações entre duas estruturas depende diretamente do mapeamento feito entre os termos comparados. Para Maedche e Staab, mapear termos de dois ou mais documentos é associar conceitos equivalentes entre eles, de acordo com uma relação de similaridade (MAEDCHE, 2002). Desta forma, para avaliar corretamente o grau de similaridade entre documentos, sugere-se considerar duas perspectivas:

- Léxicas: avalia as relações entre os termos, comparando suas cadeias de caracteres;
- Semânticas: 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 do conhecimento, pois especificam explicitamente um domínio de aplicação (GRUBER, 1993). Ao relacionar ontologias com documentos XML, permitem-se identificar o quão similares estes documentos são de determinado domínio, definindo qual ontologia melhor descreve um documento XML.

Existem algumas abordagens para análise de similaridade de documentos, como baseadas na estrutura (NIERMAN, 2002) e no conteúdo (BAEZA-YATES, 1999), de esquemas (RAHM, 2004), e de esquemas com ontologias (AUMUELLER, 2005). No entanto, o casamento de documentos XML com ontologias ainda é pouco explorado.

Existem algumas propostas para avaliação de similaridade entre esquemas que consideram tanto a perspectiva léxica quanto semântica. A integração e a implementação de abordagens bem aceitas na literatura podem gerar bons resultados para avaliação de diferentes documentos. Neste sentido, emerge a questão de pesquisa deste estudo: “É viável e eficaz a avaliação automatizada da similaridade léxica e

semântica de documentos XML e ontologias definidas em OWL, baseado na integração de propostas já aceitas na comunidade científica?".

Para responder a questão de pesquisa, definiu-se como objetivo do estudo a especificação um processo que avalie o grau de similaridade de todos os elementos definidos na representação XML e na ontologia definida em OWL, considerando tanto a perspectiva léxica quanto a perspectiva semântica. Como prova de conceito, foi desenvolvida uma ferramenta que automatiza esta avaliação, possibilitando definir qual a ontologia que melhor descreve um dado documento XML. Por fim, para avaliar os resultados obtidos pela aplicação da proposta e da ferramenta, este trabalho apresentará um estudo de caso. A principal contribuição deste trabalho é automatizar esta avaliação, possibilitando definir qual modelagem conceitual (ontologia) melhor descreve um documento XML.

Os objetivos definidos neste trabalho incluem:

- Avaliar propostas encontradas na literatura para análise de similaridade entre esquemas, visando identificar padrões e melhores práticas;
- Integrar os padrões e as boas práticas identificados, avaliando sua viabilidade prática;
- Avaliar os resultados obtidos com a proposta através de um estudo de caso.

Este trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica através de uma revisão sobre XML, ontologia e propostas de avaliação de similaridade necessárias ao entendimento do trabalho. Neste capítulo, é apresentado um estudo sobre definições de ontologias e a estrutura da linguagem de representação do conhecimento OWL (*Web Ontology Language*), sugerida pela (W3C, 2007). Além disso, são apresentadas algumas propostas para avaliação de similaridade nas perspectivas léxicas e semânticas.

O Capítulo 3 apresenta a proposta análise de similaridade entre documentos XML e ontologias definidas em OWL. Inicialmente, é sugerida a integração das abordagens estudadas para uma análise léxica e, em seguida, a análise semântica. Para avaliar a viabilidade da proposta, este capítulo apresenta a modelagem e a arquitetura da ferramenta desenvolvida, incluindo explicações contextualizadas e a dinâmica do algoritmo.

O Capítulo 4 apresenta a avaliação do processo proposto através de um estudo de caso que compara um documento XML com diversas ontologias, que especificam distintos domínios, identificando os coeficientes de similaridade. São apresentadas a organização do estudo, a instrumentação e os resultados obtidos.

O Capítulo 5 conclui o trabalho com uma síntese do que foi alcançado e algumas sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A utilização de documentos XML para o intercâmbio de dados na Web vem aumentando progressivamente. Neste contexto, a necessidade de estratégias de reuso implica na evolução de técnicas de combinação de dados para extração de informação relevante. A combinação de documentos não é trivial (DING, 2002) e depende do mapeamento realizado entre as estruturas comparadas (MAEDCHE, 2002).

Para descrever determinado domínio, sugere-se a utilização de ontologias. Ao se comparar à organização taxonômica que estrutura um documento XML com um formalismo lógico que define uma ontologia, permite-se processar não apenas as estruturas léxicas que compõem os documentos, mas também extrair relações semânticas as quais modelam o domínio.

Neste contexto, será apresentada a fundamentação teórica necessária à definição de documentos XML e ontologias, além de propostas de avaliação de similaridade entre estes documentos.

2.1 Documentos XML

A linguagem XML (*Extensible Markup Language*) foi originalmente projetada para atender os desafios da publicação eletrônica em larga escala, porém vem desempenhando um importante papel na troca de dados entre sistemas de informação (W3C, 2007).

A linguagem XML é amplamente utilizada para o intercâmbio de dados. No mundo real, bases de dados e sistemas computacionais contêm informações em formatos incompatíveis (W3SCOOL, 2007). Um dos grandes desafios enfrentados atualmente consiste na interoperabilidade destes dados.

A definição das *tags* em XML não é pré-definida, o que permite que cada sistema possa definir uma organização singular que descreva seus dados. Esta característica promove a heterogeneidade de sua estrutura em que, mesmo lidando com conceitos similares, não é trivial o processamento de sua equivalência.

Para avaliar a similaridade de documentos de um mesmo domínio, é interessante compará-los com uma estrutura que descreve este específico domínio. Ontologias são úteis neste sentido, pois definem os termos usados para descrever e representar um domínio de informação.

2.2 Ontologia

O termo "ontologia" surgiu na filosofia através de Aristóteles e significa "uma explicação sistemática da existência", isto é, a definição de um domínio do conhecimento em um nível genérico, utilizada para especificar o que existe ou o que se pode dizer sobre o mundo.

Em Ciência da Computação, ontologias representam à aquisição do conhecimento a partir de dados semi-estruturados, utilizando um conjunto de métodos, técnicas ou processos automáticos ou semi-automáticos. Dentro de Ciência da Computação, o termo "ontologia" teve sua origem na comunidade de Inteligência Artificial.

Gruber define ontologias como uma "especificação explícita de uma conceituação" (GRUBER, 1993). Uma conceituação é uma abstração simplificada da realidade que se deseja representar, isto é, um conjunto de objetos, restrições, relacionamentos e entidades que se assumem necessárias em alguma área de aplicação.

A conceituação de Gruber foi modificada por Borst, definindo ontologias como uma "especificação formal de uma conceituação compartilhada" (BORST, 1997). Esta definição enfatiza o fato que deve haver um acordo na conceituação do que é especificado.

Para se descrever uma ontologia, Maedche propõe uma estrutura (O) composta da quíntupla: O: = {C, R, Hc, rel, Ao}, onde (MAEDCHE, 02):

- C e R são dois conjuntos disjuntos formados por conceitos e relacionamentos, respectivamente;
- Hc representa a taxonomia da ontologia, isto é, a hierarquia dos conceitos e relacionamentos;
- rel representa os conceitos não taxonômicos;
- Ao representa o conjunto de axiomas da ontologia.

As ontologias vêm sendo utilizadas para descrever artefatos com diferentes níveis de estruturas (W3C, 2007). Estes níveis variam desde simples taxonomias, esquemas para meta-dados até teorias lógicas. Normalmente, elas são expressas em linguagens lógicas para que possam ser consistentes o suficiente para a extração do conhecimento.

A linguagem OWL (*Web Ontology Language*) é parcialmente mapeada através de lógica descritiva, que é um subconjunto da lógica de predicados, tornando possível um eficiente apoio lógico. Por apoio lógico, define-se o processamento do conteúdo da informação ao invés de sua simples apresentação.

2.2.1 Web Ontology Language (OWL)

Para se definir e manipular ontologias, sugere-se a utilização de linguagens que suportem estruturas para representação do conhecimento. Esta representação é realizada através da descrição formal de um conjunto de termos sobre um domínio específico. A definição de uma linguagem é necessária para a representação e descrição formal da estrutura que especifica uma conceituação.

A linguagem OWL foi recomendada pela W3C em fevereiro de 2004 como linguagem para manipulação de ontologias e o seu diferencial é a capacidade de processamento semântico através de inferência. Para se estruturar um documento OWL, que é definido sobre XML, define-se em alto nível:

- Classes: conjunto de instâncias com características comuns.
    - *Superclasse*: como as classes podem ser organizadas hierarquicamente, as instâncias diretas das subclasses são também instâncias das superclasses.
    - *Relacionamentos*: as classes podem ser sobrepostas arbitrariamente.
    - *Disjunções*: todas as classes podem potencialmente se sobrepor, porém em muitos casos é necessário fazer com que estas classes não compartilhem instâncias.
- Propriedades:
    - *Tipos* (*datatype properties*): identificam os valores primitivos das instâncias, como *integer, float, string, boolean*, etc.
    - *Objetos* (*object properties*): representam o vínculo de duas instâncias, isto é, seus relacionamentos.
    - *Inversa* (*inverseOf*): representam um relacionamento bidirecional. Adicionando valores a uma propriedade, consequentemente, se adicionam valores a uma segunda.
    - *Transitivas* (*TransitiveProperty*): se a instância x está relaciona com a instância y, e a instância y está relacionada à instância z, então x está relacionado com z. Usado principalmente em relações "parte-de".
- Indivíduos: representam os objetos em um domínio, isto é, instâncias específicas. Verifica-se aqui que dois nomes podem representar o mesmo objeto no mundo real.

A estruturação da linguagem OWL é pertinente no que tange a forma com a qual o conhecimento pode ser representado. Ao se comparar um documento XML com uma ontologia OWL, é necessário saber a equivalência dos conceitos. Para exemplificar, o relacionamento de dois recursos XML pode ser mapeado em relações taxonômicas, como *generalização* ou *especialização*, ou relações não taxonômicas, como propriedades *tipos* ou *dados,* definidas em qualquer uma das relações suportadas pela linguagem OWL.

2.3 Similaridade entre documentos

A combinação de esquemas vem sendo alvo de constante pesquisa, relacionando diferentes áreas como Banco de Dados, Sistemas de Informação e Inteligência Artificial. Podem-se destacar abordagens de mapeamento como o *Edit Distance* (LEVENSHTEIN, 1966), *Stemming* (STEMMING, 2007), *Cupid (*MADHAVAN, 2001) e *Taxonomic Overlap* (MAEDCHE, 2002).

2.3.1 Edit Distance

Para a análise léxica, o algoritmo de *Edit Distance* (ED) avalia duas cadeias de caracteres. Para esta avaliação, considera-se o número mínimo de operações necessárias para transformar uma cadeia de caractere em outra.

Por exemplo, dadas às cadeias de caracteres "*computer*" e "*computing*", a ED(*computer*, *computing*) será igual a 5, pois, para transformar a cadeia "*computer*" na cadeia "*computing*", são necessárias duas operações de remoção (respectivamente, dos

caracteres “e” e “r”) e três operações de inserção (respectivamente, dos caracteres “i”, “n” e “g”).

2.3.2 Stemming

O algoritmo de *Stemming* avalia a seqüência de caractere de determinada palavra pela remoção de seus afixos, considerando tanto prefixos quanto sufixos. Esta remoção reduz a palavra ao seu *stem*. O *stem* não é necessariamente a raiz lingüística ou radical, mas denota uma forma mínima, preferencialmente não ambígua, do termo (CHAVES, 2003).

Para exemplificar, a redução da palavra “*computer*” ou “*computing*” aos seus respectivos *stems* equivale à mesma cadeia de caractere “*comput*”.

2.3.3 Cupid

Para o mapeamento lingüístico, a abordagem *Cupid* apresenta um algoritmo genérico para combinação de esquemas genéricos, tais como XML e ontologias. O algoritmo de combinação consiste em três fases:

1. Normalização: elementos semanticamente equivalentes podem ter nomes diferentes em esquemas distintos. Neste passo, sugere-se utilizar um *Tesauro*[1] 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 lingüístico, computado entre os elementos em suas respectivas categorias.

O resultado é um valor entre Zero e Um, onde Um corresponde à boa combinação e Zero a má combinação.

2.3.4 Taxonomic Overlap

Esta comparação não avalia o elemento individualmente, mas sim o contexto em que este se encontra com relação aos demais. Maedche e Staab propõem a avaliação de similaridade entre esquemas sob duas perspectivas (MAEDCHE, 2002):

- Léxica: consideram-se as palavras que constituem os elementos;
- Semântica: considera-se o significado e a organização destes elementos.

Para a análise léxica, os autores sugerem o algoritmo *Edit Distance.* Para a análise semântica, sugere-se o conceito de *Semantic Cotopy* (SC), o qual analisa as relações hierárquicas de subconceito e superconceito do conceito em questão, formando um conjunto de conceitos pertencentes a sua hierarquia.

Para a definição do grau de similaridade entre dois conjuntos de elementos, sugere-se aplicar 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. A partir dos resultados obtidos, obtém-

[1] *Tesauro* é uma lista de palavras com significados semelhantes, dentro de um específico domínio de conhecimento.

se um valor intermediário de zero a um, similar ao Cupid, indicando o grau de similaridade.

2.4 Considerações

Este capítulo apresentou uma definição genérica de documentos XML, ontologias e propostas para avaliação de similaridades entre estes documentos. O objetivo é avaliar como conteúdos genéricos, expressos em XML, podem ser combinados com uma modelagem conceitual, expressa através de uma linguagem lógica como OWL.

Para estruturar o conhecimento definido em uma ontologia, é necessária a análise de linguagens lógicas. Para tanto, foi apresentado um estudo sobre OWL, visando analisar a forma na qual o conhecimento pode ser organizado em uma linguagem declarativa. A importância desta análise é válida para identificar as estruturas lógicas que mapeiam um modelo conceitual.

Ao mapear documentos XML e OWL, que descrevem domínios distintos, é interessante analisar diferentes sistemáticas para comparação. O objetivo deste estudo é identificar padrões e boas práticas que apóiem a aquisição do grau de similaridade entre estes documentos.

3 PROPOSTA DE AVALIAÇÃO DE SIMILARIDADE ENTRE DOCUMENTOS XML E ONTOLOGIAS

Este capítulo apresenta uma proposta de integração das diversas abordagens estudadas. O objetivo é identificar oportunidades que permitam uma cooperação entre elas, possibilitando a análise do grau de similaridade entre diferentes documentos. Com esta proposta, contribui-se com a automatização da avaliação de similaridade entre documentos XML e determinada modelagem conceitual, expressa através de uma ontologia. Por fim, este capítulo apresenta a ferramenta desenvolvida para automatizar a proposta especificada.

Um artigo chamado "*Uma proposta para análise de similaridade entre documentos XML e ontologias em OWL*" (NOLL, 2007) foi publicado nos anais da I Sessão de Pôsteres do Simpósio Brasileiro de Banco de Dados (SBBD 2007), descrevendo genericamente a proposta aqui apresentada.

3.1 Análise de similaridade léxica

Conforme apresentado no capítulo 2, existem duas abordagens principais para avaliação léxica: *Edit Distance* e *Stemming*.

Em (KANTROWITZ, 2000) é apresentado um comparativo entre algoritmos de análise léxica, incluindo *Edit Distance* e *Stemming*. 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 definidos em um XML e OWL é definida *a priori*, evitando-se erros gramaticais, sugere-se a utilização do algoritmo de *Stemming* para avaliação de similaridade léxica.

3.2 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* representa a maior base de dados léxica da língua inglesa e relaciona substantivos, verbos, adjetivos e advérbios (WordNet, 2007). Estes elementos estão agrupados em um conjunto de sinônimos cognitivos e relacionados sob a perspectiva léxica e semântica.

Outro procedimento interessante é o *Taxonomic Overlap*, que corresponde à comparação taxonômica entre os elementos avaliados (MAEDCHE, 2002). Para a definição do grau de similaridade entre dois conjuntos de elementos, sugere-se aplicar o

CJ. Para exemplificar este coeficiente, a Figura 1 apresenta um cenário onde duas árvores são avaliadas.

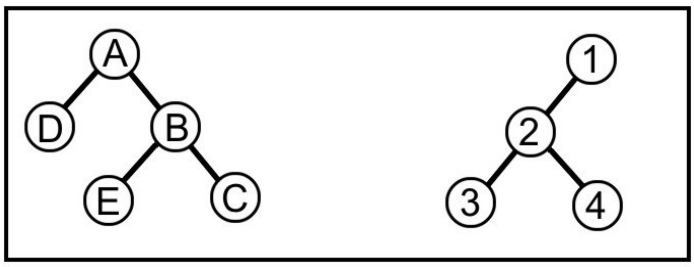

Figura 3.1: Cenário para avaliar a similaridade entre elementos

Supondo 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 corresponde ao valor três (elementos A, B e C) e que a intersecção corresponde ao valor 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. O resultado final é um coeficiente entre os valores zero e um, onde um representa uma combinação perfeita e zero representa uma má combinação.

Com o objetivo integrar as diversas propostas aqui apresentadas e avaliar sua eficácia, foi desenvolvida a ferramenta chamada *The Matcher – Uma ferramenta para avaliação de similaridade*.

3.3 *The Matcher*: Uma ferramenta para avaliação de similaridade

A ferramenta *The Matcher* comprova a viabilidade da proposta de avaliação da similaridade léxica e semântica entre documentos XML e ontologias. O objetivo desta ferramenta é verificar, dentre um conjunto de ontologias, qual a que melhor descreve um documento XML. A Figura 3.2 ilustra a interface da ferramenta desenvolvida.

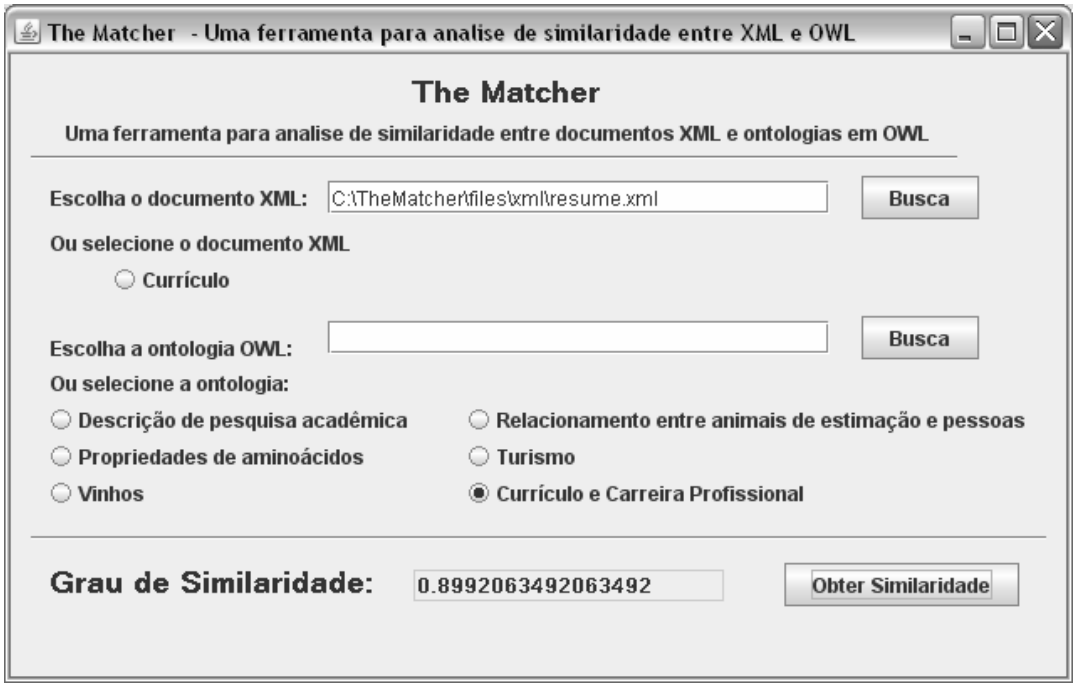

Figura 3.2: Interface da ferramenta *The Matcher*

A primeira etapa corresponde à normalização e categorização. Para cada documento (XML e OWL), foi realizado um mapeamento de todos os elementos que os compõem. Percorre-se o documento e se armazena, para cada elemento, seu radical (chave) e uma lista com o nome completo do elemento. Caso o elemento seja uma palavra composta, a lista também compreende cada um de seus componentes. Este mapeamento preliminar corresponde à perspectiva léxica e é exemplificado na Figura 3.3.

Chave	**Lista**		
Skill	SkillArea	Skill	Area

Figura 3.3: Normalização léxica dos elementos pertencentes ao XML

A próxima etapa compreende a expansão da categorização pelos seus sinônimos, isto é, adicionar à lista todas as palavras sinônimas. A relação de sinônimos é recuperada através da *WordNet* e a estrutura de dados resultante é exemplificada na Figura 3.4.

Chave	**Lista**					
Skill	SkillArea	Skill	Area	Accomplishment	Acquisition	domain

Figura 3.4: Normalização semântica dos elementos pertencentes ao XML

Esta normalização ocorre em duas etapas. Primeiramente, normalizam-se todos os recursos da ontologia, conforme descrito acima. Após, percorre-se todos os elementos do XML, verificando em sua lista a existência de alguma correspondência léxica com qualquer outro elemento pertencente a qualquer lista OWL. Todas estas correspondências são avaliadas exclusivamente pelos radicais, usando *Stemming*. Ao final deste processo, têm-se duas listas normalizadas e categorizadas, uma para OWL e outra para XML. Adicionalmente, sabe-se da existência ou não de correspondência

entre os elementos do documento XML com os recursos da ontologia, propiciando o primeiro nível de comparação entre os elementos.

A próxima etapa consiste na avaliação do *Taxonomic Overlap*, onde não se avaliam individualmente os elementos, mas a organização estabelecida desde a raiz do documento XML a cada uma de suas folhas (conjunto raiz-folha). 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 entre os elementos XML e os recursos da ontologia. Por relações na ontologia, entendem-se:

- Uma classe OWL que se relaciona com uma subclasse OWL;
- Uma classe OWL que se relaciona com propriedades *Tipo* ou *Objetos*;
- O relacionamento entre classes OWL com indivíduos quaisquer.

Para exemplificar, suponha que o conjunto A, B e C, apresentado na Figura 3.1, defina um conjunto raiz-folha de uma estrutura XML. Este conjunto possui correspondência com a árvore da direita, que representa uma ontologia. Neste exemplo, assume-se que:

- Existe correlação léxica entre os elementos A e 1 e os elementos B e 3, definida pelos radicais de algum dos elementos presentes em suas listas normalizadas;
- O elo que relaciona os elementos 1 e 2 possui uma propriedade transitiva, da mesma forma que o elo que relaciona 2 e 3, isto é, existe um elo implícito relacionando 1 e 3, através das duas propriedades transitivas.

Nesse cenário, o método de *Taxonomic Overlap* percorre a árvore da esquerda e avalia apenas os nodos que possuem correspondência com a árvore da direita, isto é, é feita uma verificação para identificar se os elementos que são lexicamente similares a A e B na ontologia possuem equivalência semântica.

Em termos práticos, como A e B se relacionam diretamente, verifica-se se seus respectivos equivalentes (1 e 3) também possuem este relacionamento. Pela propriedade transitiva descrita, consegue-se identificar que 1 se relaciona com 3, então o valor obtido pela intersecção é dois. Como o número total de elementos que se está avaliando é três (A, B e C), o resultado do CJ é dois dividido por três, isto é, 0,667.

Este procedimento é repetido para todos os conjuntos raiz-folha, completando uma lista com todos os valores obtidos pelo algoritmo. O resultado final da avaliação se dá pela média aritmética desta lista.

Para uma melhor compreensão de como o algoritmo foi estruturado, desenvolveu-se o diagrama de seqüência apresentado na Figura 3.5.

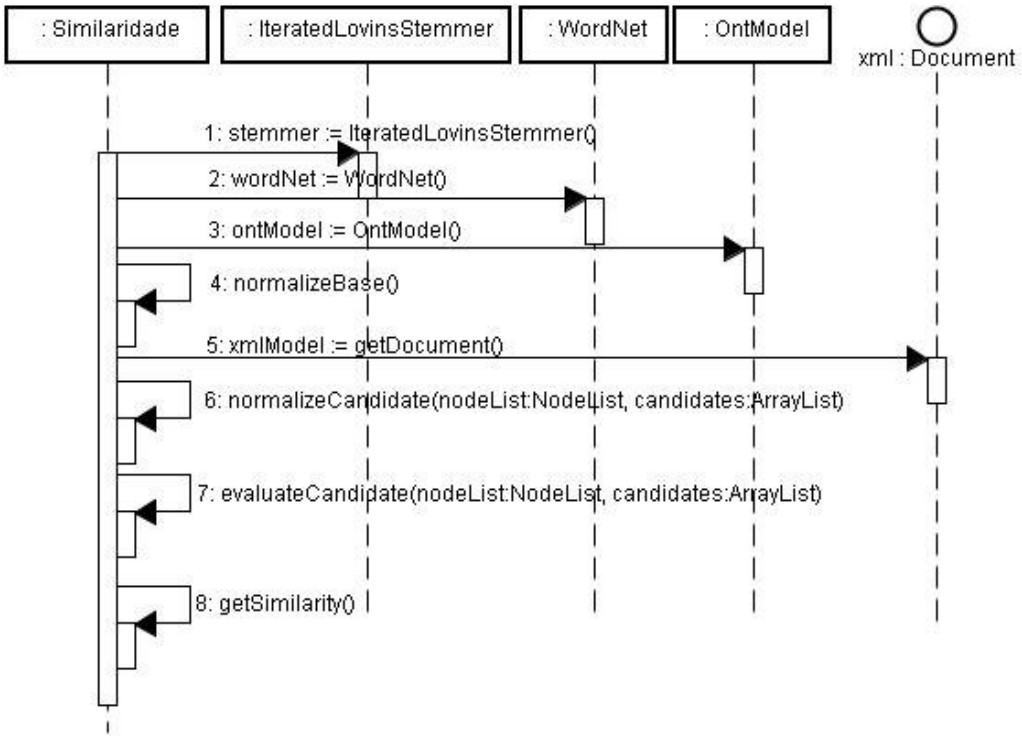

Figura 3.5: Diagrama de Seqüência para o cálculo de similaridade

A primeira interação corresponde à instanciação de duas classes auxiliares: *Stemmer* (Passo 1) e *WordNet* (Passo 2). A etapa seguinte corresponde a leitura da ontologia e sua normalização (Passo 3) em uma estrutura do tipo *HashMap*, onde a chave é o radical da palavra e o objeto é um *ArrayList* com todas as palavras da ontologia que contém este radical (Figura 3.3). Aqui se inclui as palavras sinônimas das palavras com o mesmo radical (Figura 3.4 – Passo 4).

Em seguida, ocorre a leitura (Passo 5) e normalização (Passo 6) do XML em um *HashMap,* semelhante a ontologia. Adicionalmente, o objeto do candidato contém uma referência a base (recurso da ontologia) e uma variável booleana para saber se existe correspondência entre o elemento XML e algum recurso da ontologia.

Para a avaliação de todas as relações entre os documentos, considera-se individualmente cada elemento, sem considerar a hierarquia, através de seus radicais e dos radicais de seus sinônimos (Figura 3.4 – Passo 7). Após preencher todo o *HashMap* de candidatos, ele realiza o *Taxonomic Overlap,* comparando as relações hierarquias do XML com correspondências na ontologia (Passo 8).

Para viabilizar a solução, desenvolveu-se a ferramenta utilizando a linguagem de programação Java e as seguintes API's:

- (STEMMING, 2007): necessária para avaliação léxica através da recuperação dos radicais das palavras;
- (WORDNET, 2007): necessária para avaliação semântica através da recuperação de sinônimos;
- (JWNL, 2007): necessária para acessar o dicionário relacional estabelecido pela WordNet;
- (JENA, 2007): necessária para manipulação de documentos OWL;
- (XERCES, 2007): necessária para manipulação de documentos XML.

A ferramenta *The Matcher* está disponível para download, juntamente com seu código fonte, na URL http://www.divshare.com/download/1524091-ab7.

3.4 Considerações

Como prova de conceito da proposta apresentada, foi desenvolvida a ferramenta *The Matcher*. O presente capítulo apresentou em detalhes sua sistemática, englobando uma perspectiva funcional e arquitetural. Inicialmente, apresentou-se a lógica do algoritmo e a relação de cada módulo com as propostas estudadas. Em seguida, foi apresentada a dinâmica do algoritmo através da apresentação do fluxo das atividades relacionadas.

Para comprovar a eficiência e eficácia da proposta, foi realizado um estudo de caso, apresentado no próximo capítulo. O objetivo deste estudo é avaliar se os resultados obtidos utilizando a ferramenta correspondem com a expectativa da proposta.

4 AVALIAÇÃO DA PROPOSTA

Este capítulo apresentará um estudo de caso e a avaliação dos resultados obtidos pela sua aplicação. A presente proposta de análise de similaridade apresenta duas perspectivas para sua avaliação. A primeira está relacionada com a necessidade de desenvolver um produto de software que comprove a sua viabilidade, apresentada no capítulo anterior. A segunda perspectiva está relacionada com a aplicação da proposta e seus benefícios.

4.1 Estudo de Caso

Para a definição do presente estudo de caso, optou-se por utilizar a abordagem *Goal-Question-Metric* (GQM) (BASILI, 2004). Esta abordagem parte da definição dos objetivos (nível conceitual) para o estabelecimento de questões (nível operacional), que buscam caracterizar o processo em termos de métricas (nível quantitativo):

Para a definição do objetivo global do estudo de caso, sugere-se a avaliação do 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.

A escolha das variáveis para o estudo de caso não é uma tarefa trivial. É necessário que as variáveis independentes (entrada do estudo de caso) exerçam alguma influência sobre as variáveis dependentes (saída do estudo de caso). As variáveis independentes são os documentos XML e OWL. Como variável dependente, definiu-se o grau de similaridade entre dois documentos.

A instrumentação em um estudo de caso representa a sua implementação prática, fornecendo os meios para executar este estudo. A instrumentação foi obtida a partir de duas fontes:

- Protégé para aquisição de ontologias (Protégé Library, 2007);
- XML Résumé Library para aquisição de documentos XML que descrevem currículos (XML Résumé Library, 2007).

O objetivo desta escolha é a utilização de documentos desenvolvidos sem a intervenção do pesquisador. As ontologias utilizadas no estudo 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.

Após a execução do estudo, obtiveram-se os resultados apresentados na Figura 4.6. O gráfico apresenta o grau de similaridade (eixo Y) de cada ontologia (eixo X) quando comparadas a um documento XML que define um currículo profissional.

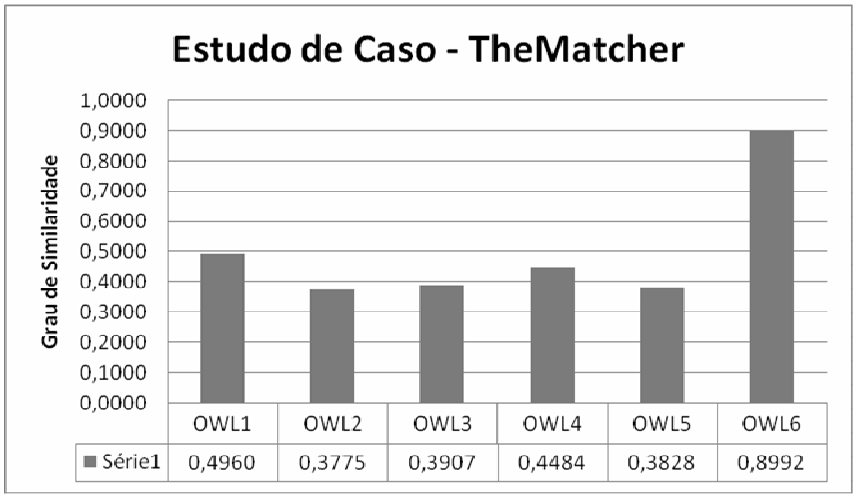

Figura 4.6: Resultados obtidos pela ferramenta.

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 ao grau obtido a partir de documentos de domínios diferentes, comprovando a eficácia e eficiência da proposta e da ferramenta.

4.2 Considerações

Este capítulo apresentou uma avaliação quantitativa da proposta de análise de similaridade léxica e semântica entre documentos XML e OWL, utilizando um estudo caso. O objetivo foi verificar a aplicabilidade e a relevância da proposta. Observou-se que os resultados obtidos através da análise de diferentes documentos, provenientes de diferentes fontes, definiu de um coeficiente de similaridade coerente com a realidade de seus conteúdos.

5 CONCLUSÕES

Este trabalho apresentou uma proposta para análise de similaridade entre documentos XML e ontologias OWL, adotando boas práticas definidas na literatura. Esta proposta sugere a integração de diversas abordagens aceitas na comunidade científica em uma sistemática funcional e bem definida.

Foi verificada a viabilidade da proposta através da modelagem e implementação de uma ferramenta. Posteriormente, foram caracterizadas a eficiência e eficácia da proposta através de um estudo de caso.

A proposta de análise de similaridade entre documentos XML e ontologias em OWL foi publicada na I Sessão de Pôsteres do XXII Simpósio Brasileiro de Banco de Dados, realizado em João Pessoa (PB), Brasil em outubro de 2007, sob o título de "Uma proposta para análise de similaridade entre documentos XML e ontologias em OWL" (NOLL, 2007).

Com o objetivo de incentivar trabalhos futuros, sugere-se:

- Extensão da base de comparação para outros documentos definidos pela (W3C, 2007) e baseados em XML;
- Variação do domínio dos documentos e das ontologias testadas.

Adicionalmente, pretende-se incorporar a ferramenta *The Matcher* 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* visam auxiliar na etapa de descoberta do domínio de conhecimento (ontologia) que descreve um conjunto de documentos XML.

REFERÊNCIAS

AUMUELLER, D.; HONG-HAI, D. Schema and ontology matching with COMA++. In: INTERNATIONAL CONFERENCE ON MANAGEMENT DATA, 25., 2005. **Proceedings…** Maryland: ACM, 2005. p. 906-908.

BAEZA-YATES, R.A.; RIBEIRO-NETO, B.A. **Modern Information Retrieval**. New York: ACM Press / Addison-Wesley, 1999.

BASILI, V.; CALDIERA, G.; ROMBACH. The Goal Question Metric Approach. In: MARCINIAC, J.J. **Encyclopedia of Software Engineering**. New York: Wiley, 1994.

BORST, W. N. **Construction of Engineering Ontologies**. 1997. 215 p. Tese (doutorado) – University of Twente, Enschede.

CHAVES, M. S. **Mapeamento e comparação de similaridade entre estruturas ontológicas**. 2003. 93p. Dissertação (Mestrado em Ciência da Computação) - PUCRS, Porto Alegre.

DING, Y.; FOO, S. A review of ontology mapping and evolving. **Journal of Information Science**, [S.1.], v.28, n.5, p.375-388, Oct. 2002.

GRUBER, T.R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal of Human and Computer Studies**, [S.1.], v.43, n.5/6, p. 907-928, 1993.

JENA: Jena Semantic Web Framework. Disponível em: < http://jena.sourceforge.net >. Acesso em: maio 2007.

JWNL: API for Java WordNet Library. Disponível em: < http://jwordnet.sourceforge.net >. Acesso em: maio 2007.

KANTROWITZ, M.; MOHIT, B.; MITTAL, V. Stemming and its effects on TFIDF Ranking (poster session). In: ANNUAL INTERNATIONAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL SIGIR, 23., 2000. **Proceedings…** Athens: ACM, 2000. p. 357-359.

LEVENSHTEIN, V. Binary Codes capable of correcting deletions, insertions, and reversals. **Cybernetics and Control Theory**, [S.1.], v.10, n.8, p.707–710, 1966.

MADHAVAN, J.; BERNSTEIN, P. A.; RAHM, E. Generic schema matching using Cupid. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES,27., 2001. **Proceedings…** New York: ACM, 2001. p. 48-58.

MAEDCHE, A.; STAAB, S. Measuring similarity between ontologies. In: EUROPEAN CONFERENCE ON KNOWLEDGE ACQUISITION AND MANAGEMENT, 2002. **Proceedings…** Madrid: ACM, 2002. p.251-263.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Massachusetts: MIT Press, 1999. 620 p.

NIERMAN, A.; JAGADISH, H.V. Evaluating Structural Similarity in XML Documents. In: INT'L WORKSHOP ON THE WEB AND DATABASES, WEBDB, 5., 2002. **Proceedings**… Madison: ACM, 2002. p. 165–176.

NOLL, R. P.; SACCOL, D. B.; EDELWEISS, N. Uma proposta para análise de similaridade entre documentos XML e ontologias em OWL. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2007. **Proceedings...** [S.l.: s.n.], 2007.

PROTEGÉ: Protegé Ontologies Library. Disponível em: < http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary >. Acesso em: maio 2007.

RAHM E.; BERNSTEIN, P.A. A survey of approaches to automatic schema matching. **The VLDB Journal: The International Journal on Very Large Data Bases**, Heildelberg, v.10, n.4, p.334-350, Dec. 2001.

SACCOL, D. B.; EDELWEISS, N.; GALANTE, R.M.; ZANIOLO, C. Managing XML Versions and Replicas in a P2P Context. In: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, SEKE, 19., 2007. **Proceedings…** Boston: ACM, 2007. p. 680-685, 2007.

STEMMING: The Lancaster Stemming Algorithm. Disponível em: < http://www.comp.lancs.ac.uk/computing/research/stemming >. Acesso em: maio 2007.

W3C: World Wide Web Consortium. Disponível em: < http://www.w3.org > Acesso em: maio 2007.

W3SCOOL: World 3 School. Disponível em: < http://www.w3school.org >. Acesso em: julho 2007.

WORDNET: A lexical database for the English language. Disponível em: < http://wordnet.princeton.edu >. Acesso em: maio 2007.

XERCES: Java Parser. Disponível em: < http://xerces.apache.org >. Acesso em: maio 2007.

XML RÉSUMÉ LIBRARY. Disponível em: < http://xmlresume.sourceforge.net >. Acesso em: maio 2007.