OWLSum: Uma Ferramenta para Sumarização de Esquemas Representados por Ontologias

Paulo Orlando Sousa[1], Carlos Eduardo Pires[2], Ana Carolina Salgado[1]

[1]Centro de Informática, Universidade Federal de Pernambuco
Av.Professor Luís Freire, s/n, Cidade Universitária – 50740-540 – Recife – PE - Brasil

[2]Departamento de Sistemas e Computação, Universidade Federal de Campina Grande
Av. Aprígio Veloso, 882, Bodocongó – 58109-970 – Campina Grande – PB – Brazil

{povqs,acs}@cin.ufpe.br, cesp@dsc.ufcg.edu.br

Abstract. *To restrict the information and highlight the important parts of content are characteristics of summaries that can be applied to ontologies. The tool OWLSum proposes to generate a summary of OWL ontologies automatically through parameters, thereby promoting a quick understanding of an ontology restricted to important concepts.*

Resumo. *Restringir a informação e destacar as partes importantes do conteúdo são características de resumos que podem ser aplicadas a ontologias. A ferramenta OWLSum propõe gerar um resumo de ontologias OWL de forma automática com o uso de parâmetros, promovendo, portanto, uma rápida compreensão de uma ontologia restrita a conceitos importantes.*

1. Introdução

A representação sucinta do conteúdo de uma fonte de dados relacional tem sido utilizada para solucionar problemas de compreensão em bancos de dados complexos [Yang and Procopiuc 2009] [Yu and Jagadish 2006]. Nesse aspecto, a sumarização de informação em estruturas de banco de dados tem mostrado bons resultados para auxiliar os usuários no entendimento de esquemas complexos de dados. Uma representação resumida dos dados promove uma rápida compreensão do que pode ser útil para diferentes formatos de dados e diversos contextos, que abrange desde a sumarização do conhecimento de um domínio, representado por uma ontologia, até o auxílio de Sistemas de Gerenciamento de Dados *Peer-to-Peer* (*Peer Data Management Systems* – PDMS).

A sumarização do conhecimento ontológico pode ser útil em diversas áreas, como, por exemplo, biologia e estudo dos genes. Na área biológica, a sumarização pode auxiliar no estudo genético, que utiliza ontologias para mapear genes [Lin and Sakamoto 2009]. Em pesquisas, realizadas na área genética, verifica-se a similaridade entre vários genes. A sumarização permite otimizar as pesquisas, pois, ao invés de utilizar todo o conteúdo da ontologia, apenas a parte de maior importância (subontologia) é explorada na comparação dos genes.

No caso do PDMS proposto por [Pires 2009], a sumarização ganha destaque para auxiliar na organização (*clustering*) de *peers* em uma rede P2P. A arquitetura do sistema agrupa os *peers* em *clusters*, organizando-os semanticamente de acordo com o esquema dos mesmos. Cada *cluster* de *peers* contém uma representação esquemática dos *peers* que o compõe, o que facilita a inclusão de novos *peers*. Essa representação simplifica o processo de inclusão de novos *peers* no sistema, pois irá comparar o esquema (ontologia) do novo *peer* apenas com o sumário da ontologia que representa o *cluster*. De acordo com

os processos apresentados no PDMS, foram observadas duas necessidades importantes para a ferramenta de sumarização de ontologias: (i) gerar os sumários automaticamente, pois, no caso do PDMS, o esquema que representa o *cluster* será atualizado cada vez que houver a inserção ou remoção de um *peer*; e (ii) aferir a qualidade do sumário gerado, que precisa ser conciso e fiel às expressividades da ontologia original.

Para esses cenários, propomos a OWLSum, uma ferramenta automática para sumarização de ontologias. As principais contribuições da ferramenta são: (i) geração automática de resumos de ontologias com tamanhos parametrizáveis; (ii) capacidade de gerar uma subontologia considerando os critérios de centralidade e freqüência para garantir a participação dos conceitos mais relevantes no sumário; e (iii) uso de métricas de recuperação de informação para avaliar a qualidade dos sumários.

Este artigo está organizado da seguinte forma: A Seção 2 apresenta uma visão geral da ferramenta. A Seção 3 descreve as medidas de centralidade e freqüência. A Seção 4 mostra o processo de sumarização. A Seção 5 apresenta a ferramenta proposta. Finalmente, na Seção 6, concluímos o nosso trabalho e damos uma perspectiva futura.

2. Visão Geral

O processo da sumarização automática de ontologia possibilita formas de simplificar e agilizar algoritmos que usam ontologias. Assim como o objetivo de um resumo é restringir a informação e dar destaque às partes importantes do conteúdo, o resumo de uma ontologia pode diminuir o tempo de processamento dos algoritmos através da simplificação oferecida pela sumarização. Este processo gera uma ontologia sumarizada contendo a parte de maior importância da ontologia original.

As ontologias são expressas em linguagens formais, bem definidas semanticamente, como a Web Ontology Language (OWL), linguagem utilizada nessa ferramenta e padrão da W3C [McGuinness and Harmelen 2004]. A Figura 1 ilustra o processo de sumarização da aplicação que consiste em: dada uma ontologia de entrada *O* gerar uma versão resumida, chamando-a de ontologia sumarizada (denotado por *OS*). Os conceitos relevantes de *O* (representados na cor cinza) são inicialmente identificados para a geração de *OS*, que corresponde a uma subontologia de *O*, concentrando o número máximo de conceitos relevantes. Como os conceitos identificados como relevantes podem ser não-adjacentes em *O*, é possível que conceitos menos importantes (cor branca) sejam introduzidos em *OS*. Tais conceitos "não-relevantes" são necessários para manter a integridade e preservar os relacionamentos entre os conceitos relevantes da ontologia original. Por isso, *OS* corresponde a uma subontologia de *O*, contendo o mínimo de conceitos não-relevantes devidamente interconectados, evitando qualquer intervenção humana.

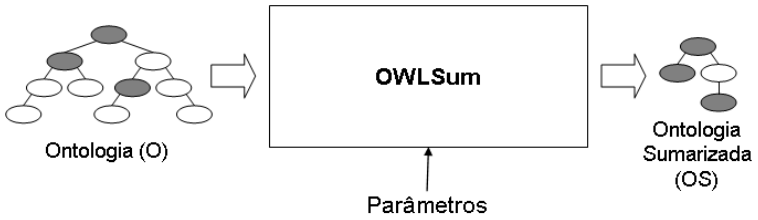

Figura 1. Uma visão geral do processo de sumarização ontológica

Na ferramenta, a ontologia *O* é modelada como um grafo direcionado com conexões rotuladas $O = (C, R)$, onde $C = (c_1, ..., c_n)$ é um conjunto finito de vértices (conceitos) e $R = (r_1, ..., r_n)$ é um conjunto finito de arestas (relacionamentos entre os conceitos). Da mesma forma, definir um resumo de uma ontologia *OS* é como criar um subgrafo de *O* no qual $OS \subset O$. Formalmente, $OS = (CS, RS)$, onde $CS \subset C$ e $RS \subset R$.

3. Medidas de Relevância

A relevância de um conceito c_n em uma ontologia O é medida considerando os relacionamentos de c_n com outros conceitos em O (centralidade) e as ocorrências de c_n nas ontologias que formaram O (freqüência). Em nossa abordagem, a centralidade é utilizada para captar a importância de um conceito dentro de uma ontologia. A freqüência é usada quando uma ontologia O (ontologia do cluster) é resultante de um processo de integração (*merging*) de ontologias $O_1,...,O_n$ (ontologias que representam os *n peers* que compõem o *cluster* [Pires 2009]) e captura as ocorrências dos conceitos nas ontologias geradoras de O.

3.1. Medida de Centralidade

Centralidade [Mika 2007] é uma das maneiras mais importantes para identificar vértices relevantes dentro de um grafo. As medidas de centralidade mais usadas são: *degree, closeness e betweenness*. A medida *degree centrality* [Mika 2007] defende a idéia de que um vértice *v* com grande número de ligações para *n* vértices proporciona uma ampla cobertura de acesso entre os vértices do grafo. Neste trabalho, estendemos a definição original da medida *degree centrality* não só para considerar o número de relacionamentos entre os conceitos da ontologia, mas também para os tipos de relacionamento entre eles. Existem dois tipos de relacionamento: *standard* e *user-defined*. Os relacionamentos *standard* são: *is-a*, *part-of* e *same-as*. Os relacionamentos *user-defined* são aqueles definidos pelo usuário e dependentes do domínio como, por exemplo, *hasItems* e *authorOf*. A fórmula de normalização definida para o grau de centralidade de um conceito c_n é:

$$centrality(c_n) = \frac{nr \times \left(\frac{n_s \times w_s}{max_s} + \frac{n_{ud} \times w_{ud}}{max_{ud}} \right)}{|C|-1}, \text{ onde } centrality(c_n) \in [0,1]$$

O número de relacionamentos em c_n obedece duas regras para contagem: ignorar os auto-relacionamentos do conceito e contabilizar um único relacionamento caso haja mais de um entre dois conceitos. Essa prática tenta atender à idéia *degree centrality* que defende o amplo acesso entre os conceitos. n_s e n_{ud} são, respectivamente, o número de relacionamentos *standard* e *user-defined* mantidos em c_n. w_s e w_{ud} são, respectivamente, os pesos para n_s e n_{ud}, na qual ($w_s + w_{ud} = 1$). max_s e max_{ud} representam, respectivamente, o número máximo de n_s e n_{ud} mantidos por um certo conceito na ontologia. n_r representa o número de relacionamentos a conceitos distintos, que c_n mantém tornado $n_s + n_{ud} = n_r$.

3.2. Medida de Freqüência

A freqüência é uma medida que pode ser usada em uma ontologia integrada O, obtida com o resultado da fusão (*merging*) entre várias ontologias $O_1,...,O_n$ [Noy and Musen 2000]. A Tabela 1 mostra exemplos de correspondências entre conceitos de O e de $O_1,...,O_n$. Por exemplo, o conceito *Faculty* da ontologia O é identificado como: (i) equivalente a *Faculty* na ontologia O_1; (ii) subconceito de *Worker* na O_2; e (iii) superconceito de *Professor* em O_3 e *PostDoc* em O_4.

Tabela 1. Exemplo de conceitos correspondentes

Correspondências para o conceito O:Faculty	
O:Faculty $\equiv$ O_1:Faculty	O:Faculty $\sqsupset$ O_3:Professor
O:Faculty $\sqsubset$ O_2:Worker	O:Faculty $\sqsupset$ O_4:PostDoc

Neste trabalho, supomos que *O* pode ser uma ontologia gerada por *merging* na qual um conceito $c_n \in C$ corresponde a um ou mais conceitos contidos em $O_1, ..., O_n$. Nesse sentido, a fórmula definida para freqüência de um conceito c_n é:

$$frequency(c_n) = \frac{|correspondences(c_n)|}{n}, \text{ onde } frequency(c_n) \in [0,1]$$

Em outras palavras, $frequency(c_n)$ é definida como a razão entre o número de conceito correspondentes que envolvam c_n (denotado $|correspondences(c_n)|$) e o número de ontologias diferentes que formaram *O* (indicado por *n*). Ambas as informações podem ser extraídas das correspondências que foram geradas no processo de *merging*.

4. Construindo um Resumo de Ontologia

As principais etapas do processo da sumarização de ontologias são: (i) calcular a relevância dos conceitos da ontologia; (ii) determinar os conceitos relevantes; (iii) agrupar os conceitos relevantes adjacentes; (iv) identificar os caminhos entre os grupos de conceitos; (v) analisar os caminhos identificados; e (vi) definir o resumo ontologia.

1) Calcular a relevância dos conceitos: nossa proposta consiste em combinar centralidade e freqüência em uma fórmula ponderada, com os pesos definidos pelo usuário, de acordo com a importância de cada medida. A seguinte fórmula é usada para calcular a pertinência de um conceito particular c_n em uma ontologia *O*:

$$relevance(c_n) = \lambda \times centrality(c_n) + \beta \times frequency(c_n), \text{ onde relevance}(c_n) \in [0,1].$$

λ e β são pesos associados, respectivamente, à centralidade e freqüência, e $\lambda + \beta = 1$.

2) Determinar os conceitos relevantes: esta etapa consiste em identificar o conjunto de conceitos relevantes (denotado RC, onde RC $\subseteq$ C) de uma ontologia *O*. Na ferramenta considera-se que RC tem um tamanho *k* definido pelo usuário para restringir os *k* conceitos mais relevantes de *O*.

3) Agrupar os conceitos relevantes adjacentes: esta etapa consiste na formação de grupos de conceitos relevantes, contendo apenas conceitos adjacentes nos grupos da ontologia inicial *O*. Na construção dos grupos de conceitos podem ocorrer as seguintes situações: (i) cada grupo se formou por um único conceito relevante (todos os conceitos relevantes são não-adjacentes em *O*); (ii) pelo menos um dos grupos tem mais de um conceito relevante (alguns conceitos relevantes são não-adjacentes em *O*); e (iii) um único grupo é formado, contendo todos os conceitos relevantes. Nas duas primeiras situações, o processo da sumarização prossegue com as etapas 4, 5 e 6. Na última situação, o processo conclui com o sumário correspondendo ao único grupo formado.

4) Identificar os caminhos entre os grupos de conceitos: se houver pelo menos dois grupos de conceitos na ontologia inicial *O* (situações i e ii da Etapa 3), todos os caminhos entre os grupos conceituais de *O* são detectados. Esta etapa enumera todos os caminhos possíveis entre cada par de grupos.

5) Analisar os caminhos identificados: múltiplos caminhos entre os grupos de conceitos podem ser identificados. As métricas clássicas de *recall* e *precision*, comumente utilizadas em Recuperação de Informação [Baeza-Yates 1999], são aplicadas para determinar o nível de cobertura e precisão de cada caminho, respectivamente. *Recall* representa o número máximo de conceitos relevantes no caminho, já *precision* representa o número mínimo de conceitos não-relevantes no caminho.

$$Recall = \frac{|Path_i \cap RC|}{|RC|} \qquad Precision = \frac{|Path_i \cap RC|}{|Path_i|}$$

Caminhos não podem ser comparados baseados somente na precisão e cobertura. Um caminho pode ter uma alta cobertura e baixa precisão, e vice-versa. Nesse caso, utilizamos a formula *f-measure* [Baeza-Yates 1999] para combinar precisão e cobertura conforme o parâmetro α, inserido pelo usuário.

$$f-measure = \frac{Precision \times Recall}{(1-\alpha) \times Precision + \alpha \times Recall}$$, onde $\alpha \in [0,1]$.

6) Determinar o resumo da ontologia: a seleção do melhor caminho candidato é feita de acordo com: (i) *f-measure*: o caminho deve ter o número máximo de conceitos relevantes e o mínimo de conceitos não-relevantes, ou seja, o caminho com o maior valor de *f-measure* deve ser selecionado; (ii) relevância média: caso haja empate no valor de *f-measure* então usar a média de relevância (razão entre a soma da relevância dos conceitos e o número de conceitos no caminho).

5. A Ferramenta OWLSum

Para demonstrar a ferramenta OWLSum utilizaremos um exemplo no qual são fornecidos como entrada: uma ontologia *O* (*Education.owl*) proveniente de um processo de *merging* e um arquivo XML contendo as correspondências entre os elementos de *O* e os elementos de 10 ontologias que formaram *O*. Na Figura 2, após abrir a ontologia *O*, tem-se a representação da mesma em formato de grafo, contendo em cada nó (conceito) o número de correspondências que cada conceito possui com os conceitos das ontologias geradoras de *O*. Por exemplo, o conceito *University* possui correspondências com conceitos de 8 ontologias geradoras.

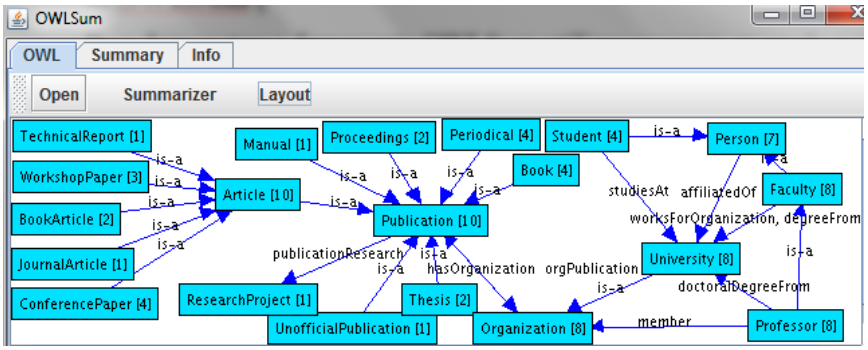

Figura 2. A ilustração da ontologia *Education.owl* na ferramenta *OWLSum*

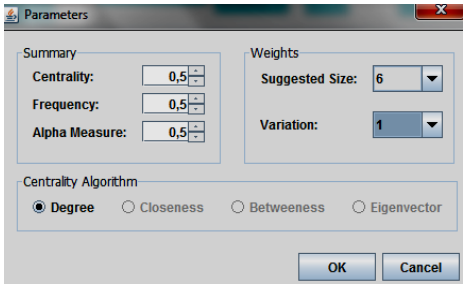

Figura 3. Janela de parametrização *do OWLSum*

Para iniciar a sumarização é preciso clicar em "Summarizer" e indicar o local no sistema de arquivos onde a ontologia sumarizada será criada. Logo depois, aparece a janela "Parameters" para receber os valores dos parâmetros. Como mostrado na Figura 3: (i) os campos "Centrality" e "Frequency", por *default*, são instanciados com valores iguais (0,5) para dar a mesma importância a *centrality* e *frequency* no cálculo de *relevance*; (ii) o campo "Alpha Measure" também é preenchido com 0,5 para dar

equivalente importância a *Recall* e *Precision* no cálculo de *f-measure*; (iii) "Suggested Size" e "Variation" possuem valores 6 e 1, respectivamente, para estabelecer o tamanho e a variação da ontologia a ser gerada. Após fornecer os parâmetros, realiza-se o processo de sumarização que irá analisar os caminhos entre grupos de conceitos (subgrafos), contendo entre 5 e 7 conceitos, e que possuam o maior número de conceitos relevantes e a menor quantidade de conceitos não-relevantes. A ontologia sumarizada (Figura 4) é gerada em um arquivo OWL no diretório especificado pelo usuário.

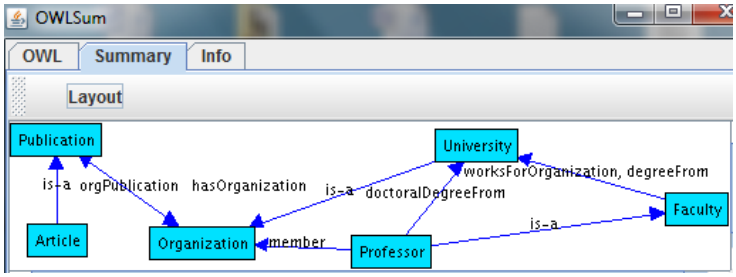

Figura 4. Sumário da ontologia *Education.owl* na ferramenta *OWLSum*

6. Conclusão

As principais contribuições da ferramenta OWLSum consistem na sumarização automática de uma ontologia OWL e na análise de qualidade do resumo gerado, através dos cálculos de *f-measure* e relevância média. O resumo apresenta grande potencialidade de uso para diversas aplicações. Como mencionado nesse trabalho, a OWLSum pode ser empregada para agilizar a compreensão das ontologias, possibilitando um processamento mais rápido de ontologias de grande porte, além de considerar a medida de freqüência para ontologias formadas a partir de outras ontologias. Portanto, a ferramenta OWLSum é bastante útil para aplicações que manipulam grandes ontologias. Em uma próxima versão, a ferramenta deverá incluir as medidas de centralidade *closeness e betweenness*, possibilitando assim uma análise mais detalhada dos resumos gerados.

Referências

Baeza-Yates, R., Ribeiro-Neto, B. (1999) "Modern Information Retrieval", Harlow, England, ACM Press.

Lin, Y., Sakamoto, N. (2009) "Ontology Driven Modeling for the Knowledge of Genetic Susceptibility to Disease", The Kobe Journal of the Medical Sciences 54(6), pp. 290-303.

McGuinness, D. L., Harmelen, F. V. (2004) "OWL Web Ontology Language Overview", http://www.w3.org/TR/2004/REC-owl-features-20040210, February.

Mika, P. (2007) "Social Networks and the Semantic Web", Springer-Verlag, New York.

Noy, N. F., Musen, M. A. (2000) "Prompt: Algorithm and Tool for Automated Ontology Merging and Alignment", in Proc. of AAAI'00, Austin, USA, pp. 450-455.

Pires, C. E. S. (2009) "Ontology-based Clustering in a Peer Data Management System", Ph.D. Thesis, Universidade Federal de Pernambuco (UFPE/CIn). Recife, PE, Brazil.

Yang, X., Procopiuc, C. M., Srivastava, D. (2009) "Summarizing Relational Databases", in Proc. of the VLDB Endowment, Volume 2 , Issue 1, pp. 634-645.

Yu, C., Jagadish, H. V. (2006) "Schema Summarization", in Proc. of VLDB'06, Seoul, Korea, pp. 319-330.