

Ontologias Autônomicas

Stainam Brandao
COPPE/UFRJ

Ilha do Fundão - CT - Lab: H-307
CEP:21.941-972 - Rio de Janeiro
Tel.: (+55 +21) 2562-8783

stainam@cos.ufrj.br

Jonice Oliveira
IM/DCC/UFRJ

Ilha do Fundão - CT - Lab: H-307
CEP:21.941-972 - Rio de Janeiro
Tel.: (+55 +21) 2562-8783

jonice@cos.ufrj.br

Jano M. Souza
COPPE/UFRJ

Ilha do Fundão - CT - Lab: H-307
CEP:21.941-972 - Rio de Janeiro
Tel.: (+55 +21) 2562-8783

jano@cos.ufrj.br

RESUMO

Documentos web anotados semanticamente possuem informações além das que estão expressas através dos textos em linguagem natural. Essas anotações descrevem o conteúdo do documento de uma forma explícita e inteligível para que máquinas ou agentes de software possam validar e assegurar a fonte de informação, como também inferir conhecimento além das expressas explicitamente. Contudo, esta não é a realidade de um ambiente independente como a Web, onde não existe nenhuma restrição na informação sendo publicada, o que pode comprometer a qualidade. Neste artigo focamos na representação do conhecimento disponível na Web com ontologias de domínio e o processamento deste conhecimento de forma autônoma através de uma sistemática de avaliação e re-projeto da própria ontologia. Este comportamento autônomo do repositório pretende garantir o auto-gerenciamento da ontologia com uma estratégia de evolução para tratar o ambiente dinâmico em que está inserido, corrigindo problemas identificados e realizando adaptações necessárias para otimizar e proteger contra eventos indesejáveis. A abordagem da computação autônoma foi utilizada para fornecer às ontologias capacidade de adaptação no ambiente Web e fornecer serviços adicionais ao usuário ou executar tarefas mais complexas.

Categorias e Descritores do Assunto

H.3.1 [Content Analysis and Indexing]: Relacionado com avaliação da ontologia usando métricas que podem ser monitoradas e disparar ações autônomas.

Termos Gerais

Gerenciamento, Medida, Projeto e Verificação.

Palavras-Chave

Web Semântica, Representação do Conhecimento, Ontologias.

1. INTRODUÇÃO

Ontologias são definidas neste trabalho como uma representação da conceituação [22] e, portanto, uma representação simbólica de conceitos do mundo, tendo uma visão epistemológica da realidade e uma possibilidade de mundo coerente [4]. No entanto, vale ressaltar a que a linguagem e a influência do senso comum sobre a construção de ontologias pode filtrar esta realidade, já que conceitos podem ter diferentes significados quando aplicado em diferentes contextos.

Ontologias provaram-se benéficas para a representação do conhecimento de domínio [6] e sua importância para a web

semântica [21]. Neste último caso, muitas vezes ontologias apóiam o processo de indexação de conteúdo dos recursos, como uma forma de anotação semântica, resultando na representação do conhecimento explícito que não pode ser avaliado e gerenciado em um ambiente independente como a Web. No entanto, as ontologias como um modelo conceitual de um domínio de negócios deve reagir a quaisquer mudanças no ambiente empresarial, sem afetar o modelo pretendido e também incorporar funcionalidades adicionais, em conformidade com as mudanças nas necessidades do usuário como, por exemplo, a evolução da estrutura taxonômica da ontologia. Se a evolução da ontologia ou a anotação semântica é executada de forma redundante, inconsistente ou incompleta, então a confiabilidade, precisão e eficiência do sistema diminuem significativamente [1]. Segundo esses autores, a fim de evitar estes problemas reais, as aplicações baseadas em ontologias devem suportar mecanismos de detecção de mudanças, analisar e resolver de uma forma consistente. Mas para isso, é necessário que as ontologias de domínio tenham acesso aos documentos Web que estão sendo anotados ou tenham acesso aos documentos que pertencem ao mesmo domínio da ontologia. Neste trabalho, apresentamos uma proposta para anotação semântica que visa a evolução da ontologia sobre a conceituação baseada em métricas de qualidade, consulta de usuários sobre a ontologia e a anotação semântica de documentos Web. Mas para isso, o primeiro passo é o mapeamento dos documentos do domínio utilizando para isso uma técnica de sumarização do conteúdo dos textos e o conhecimento do domínio representado na ontologia.

2. ARCABOUÇO AUTÔNOMO PARA REPRESENTAÇÃO DO CONHECIMENTO WEB

Como a construção manual de ontologias requer um esforço humano enorme, a aquisição delas é considerada um gargalo para a Web Semântica [2]. Além disso, existe a dificuldade para a manutenção, onde ontologias pequenas podem não ser muito representativas e as maiores podem ser complexas para impedir sua re-utilização. Dentro deste trabalho, manutenção envolve tanto ações corretivas quanto evolutivas através de ações autônomas que atuam no vocabulário (taxonomia) da ontologia que está representando documentos da web.

Abaixo, descreveremos um repositório que de forma autônoma evolui as ontologias de domínio através dos documentos Web indexados e das consultas realizadas sobre elas. Embutimos nas ontologias de domínio o autogerenciamento para tratar o dinamismo do ambiente Web através da identificação de

eventos que demandem nova configuração, correção, otimização e proteção. O monitoramento de eventos está associado a ações especificadas para reduzir a necessidade por intervenção humana. Os eventos são identificados através de métricas de avaliação de ontologias citadas na literatura. Já os padrões foram definidos dentro deste trabalho com o objetivo de utilizar as métricas como parâmetros e garantir a ação autônômica sobre a conceitualização sem afetar o modelo pretendido pelo engenheiro de ontologia. Porém, o engenheiro continua sendo o responsável por auditar as ações autônômicas, ratificando ou revogando a evolução da ontologia.

O primeiro passo para isso foi a definição das categorias de métricas que serão monitoradas pelo repositório. Qualquer mudança nos valores das métricas de uma ontologia é detectada e informada ao módulo responsável pela análise deste evento. Evento dentro deste trabalho é definido como uma mudança nas métricas de avaliação da ontologia modeladas no arcabouço. As métricas atualmente utilizadas se enquadram em 3 (três) categorias: Tamanho Intencional, Estrutura Taxonômica e Aplicação da Linguagem Natural.

Qualquer alteração nos valores das métricas é acompanhada de uma análise na configuração, proteção, cura ou otimização definida para as ontologias. Para isso, expressões de guarda são definidas para valores de métricas que disparam ações autônômicas. A figura abaixo mostra um exemplo onde a métrica *Precision* possui uma expressão de guarda relacionada à Configuração (C), enquanto a métrica *Cost-Based Evaluation* possui uma expressão de guarda que dispara uma ação de Cura (H).

Category	Metric	Value	C	H	O	P
Natural Language Application	Precision	0.0	✗			
	Recall	0.0				
	Coverage	0.0	✗			
	Accuracy	0.0	✗			
	Cost-Based Evaluation	0.0		✗		
Taxonomic Structure	Lexical Comparison Level	0.0				
	Width	1.22			✓	
	Depth	50.0			✓	
Instance Structure	Specificity	50.0			✓	
	Completeness	0.0				
	Importance	0.0				✗
	Average Population	0.0				✗
	Class Richness	0.0				

Figura 1. Exemplo de expressões de guarda disponíveis por métricas.

Uma métrica pode acionar mais de uma ação autônômica. Por isso, é necessário orquestrar as ações disparadas após um evento evitando disparos infinitos e resultados inesperados para a evolução da mesma pelo próprio repositório.

Atualmente, os padrões de ação disponíveis arcabouço são: Abstração da Ontologia, Fragmentação e Balanceamento Semântico. O desafio relacionado com a captura, modelagem, armazenamento e monitoramento de cenários justifica a necessidade de uma arquitetura apropriada para este fim. A arquitetura do sistema foi desenvolvida utilizando a abordagem da computação autônômica para prover capacidade de autogerenciamento e esconder a complexidade de programação do usuário.

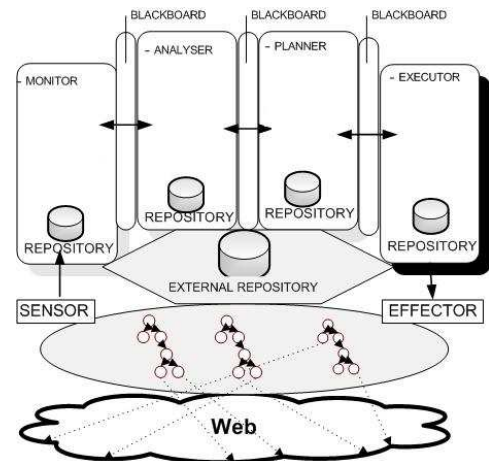


Figura 2. Arquitetura do repositório.

A figura acima apresenta a arquitetura com os quatro módulos e, abaixo uma breve descrição destes:

- *Monitor*: possui sensores que detectam eventos nas ontologias. Definimos como evento mudanças relativas às consultas realizadas, novos documentos Web associados à ontologia e manipulação da ontologia por parte do engenheiro. Quando isso acontece, o repositório recalcula as métricas e havendo alguma mudança informa ao módulo *Analysier*.
- *Analysier*: possui Regras Ativas que verificam se alguma expressão de guarda foi ativada após a ocorrência de algum evento. Toda expressão de guarda está associada a uma ação autônômica.
- *Planner*: quando várias expressões são disparadas simultaneamente, este módulo orquestra para evitar disparos infinitos e resultados inesperados.
- *Executor*: gerencia os padrões de ação especificados pelo arcabouço para que quando solicitado pelo Planejador, atue na ontologia gerando o resultado esperado com a evolução da mesma.

Para a comunicação entre os módulos utilizamos a abordagem *blackboard*, definido como um espaço de conhecimento onde entidades fracamente acopladas compartilham informação [17]. Os módulos registram no blackboard todos os fatos ocorridos com a ontologia e é responsabilidade de cada um ler a informação disponibilizada.

3. RESULTADOS PRELIMINARES

Uma ontologia do domínio da previdência social do Brasil foi desenvolvida com o objetivo inicial de representar o vocabulário e um objetivo futuro de mapear documentos do domínio, Leis e normas da previdência disponível em bibliotecas on-line. Atualmente a ontologia possui 200 conceitos e está disponibilizada no repositório implementado.

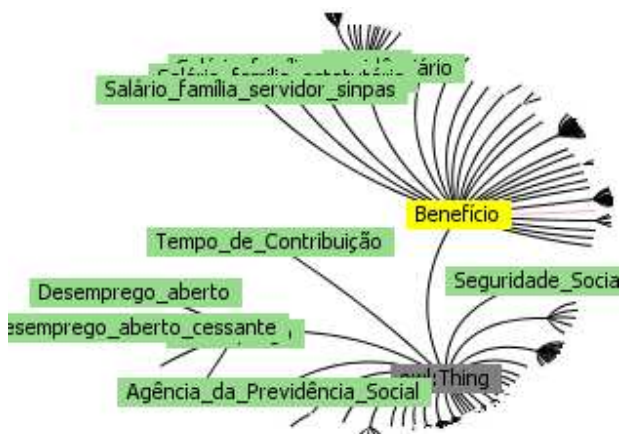


Figura 3. Captura da ontologia do domínio da Previdência Social.

3.1 Objetivo

O objetivo deste Estudo de Caso foi avaliar a abordagem apresentada neste artigo para gerar uma abstração da ontologia baseada na coleção de documentos disponibilizados pelo usuário no repositório da ontologia. Para isso, o primeiro passo foi a definição da URLs que contém a fonte de documentos, serviços on-line e o sistema de legislação social para qual será gerada uma visão da ontologia: www.previdenciasocial.gov.br/ e <http://www81.dataprev.gov.br/sislex/>, respectivamente.

3.2 Resultados

Após a execução de uma metodologia para Anotação Semântica, aproximadamente 600 documentos foram identificados e aproximadamente 400 mil palavras foram analisadas, processadas e seus documentos associados à conceitos da ontologia. O passo seguinte foi a execução da ação autônoma de Abstração para que um vocabulário seja gerado para a coleção de documentos identificados, utilizando conceitos da ontologia e a identificação de termos candidatos.

Tabela 1. Resultado obtido do Estudo de Caso

Objetivos	Valores
Conceitos da ontologia identificados	42
Termos candidatos identificados	35
Precisão	59%
Cobertura	81%
Acurácia	72%

Em seguida, o vocabulário gerado foi analisado por dois especialistas da área de Produtos de Benefício da Previdência Social, que baseado nos sumários dos documentos e os conceitos a que estes foram associados, apontaram os conceitos que deveriam e não foram associados, bem como os conceitos que não deveriam, mas foram equivocadamente associados ao documento.

Um módulo desenvolvido exclusivamente para este estudo utilizou estas informações para calcular as métricas de Precisão, Recall, Acurácia e Cobertura da ontologia diante dos documentos Web associados e já avaliados pelos participantes do Estudo de Caso (Tabela 1). Dentro deste trabalho, Precisão é considerada a relação entre o número de itens corretamente identificados e o número de itens identificados. Já o Cobertura mede a relação entre o número de itens corretamente identificados e o número total de itens corretos. E finalmente, Acurácia é a média da cobertura daquela frequência de classes que incluem pelo menos 60% das palavras relevantes.

A mesma interface permite visualizar os documentos que deram origem a identificação de cada conceito ou termo candidato do domínio (figura 4).

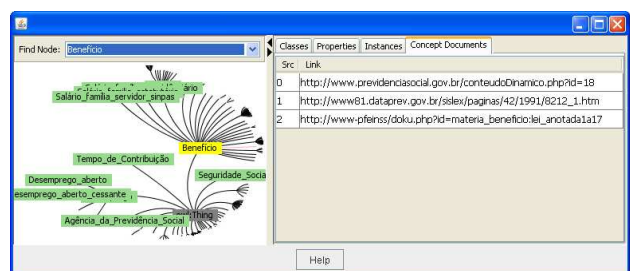


Figura 4. Documentos associados ao conceito 'Benefício'

Na mesma figura 4, temos o conceito Benefício do lado esquerdo com três links (lado direito) de onde este conceito foi identificado.

4. CONCLUSÃO E CONSIDERAÇÕES FINAIS

Apresentamos neste trabalho uma metodologia para instanciar documentos web na ontologia como uma forma de anotação semântica para representação do conhecimento explicitado na Web, como também prover à ontologia de domínio uma fonte de recursos do domínio que ela representa. A metodologia foi implementada através da ferramenta WebOnto e o domínio da previdência social do Brasil está sendo utilizado no estudo de caso. Os resultados ainda estão sendo gerados e como trabalho futuro pretendemos prover autogerenciamento às ontologias de domínio através da definição de uma estratégia de evolução para tratar o ambiente dinâmico em que está inserido. Isso será feito através da correção de problemas identificados e adaptações necessárias para otimizar e proteger contra eventos identificados, através das métricas calculadas automaticamente ou fornecidas pelo especialista de domínio.

5. REFERENCES

- [1] A. Maedche, B. Motik, L. Stojanovic, R. Studer, e R. Volz, "An infrastructure for searching, reusing and evolving distributed ontologies," WWW, 2003, pp. 439–448.
- [2] B. Omelayenko, "Learning of Ontologies for the Web: the Analysis of Existent Approaches," Proceedings of the International Workshop on Web Dynamics, held in conj. with the 8th International Conference on Database Theory (ICDT'01), London, UK, 2001.

- [3] Clips, "<http://clipsrules.sourceforge.net/>" Accessed in august 22th 2009.
- [4] G. Guizzardi, "Ontological foundations for structural conceptual models," 2005.
- [5] G.A. Miller, "WordNet: A Lexical Database for English," COMMUNICATIONS OF THE ACM, vol. 38, 1995, pp. 39--41.
- [6] H. Alani e C. Brewster, "Ontology ranking based on the analysis of concept structures," Proceedings of the 3rd international conference on Knowledge capture, Banff, Alberta, Canada: ACM, 2005, pp. 51-58.
- [7] Huebscher, M. C. and McCann, J.A., "A survey of autonomic computing—degrees, models, and applications", ACM Computing Surveys, 2008.
- [8] H. Ehrig, W. Reisig, e G. Rozenberg, Petri Net Technology for Communication-Based Systems: Advances in Petri Nets (Lecture Notes in Computer Science, 2472), SpringerVerlag, 2004.
- [9] IBM, "An architectural blueprint for autonomic computing", Autonomic Computing White Paper, Third Edition, June 2005.
- [10] Jaccard, "<http://sourceforge.net/projects/simmetrics/>" Accessed in august 22th 2009.
- [11] Jess, "<http://www.jessrules.com/>" Accessed in august 22th 2009.
- [12] Kephart, J.O., Chess, D.M. "The Vision of Autonomic Computing", IBM Thomas J. Watson Research Center, 2003.
- [13] K. Beyer, J. Goldstein, R. Ramakrishnan, e U. Shaft, "When Is "Nearest Neighbor" Meaningful?," IN INT. CONF. ON DATABASE THEORY, 1999, pp. 217--235.
- [14] Miller, B., The autonomic computing edge: Can you CHOP up autonomic computing? IBM Corporation, 2005.
- [15] M.W. Geert, G., "The Ontological Foundation of REA Enterprise Information Systems," Ago. 2000.
- [16] N. Guarino, "Formal Ontology and Information Systems," 1998.
- [17] M. Shaw e D. Garlan, Software Architecture: Perspectives on an Emerging Discipline, Prentice Hall, 1996.
- [18] N. Russell, A. ter Hofstede, D. Edmond, e W. van der Aalst, "Workflow Data Patterns: Identification, Representation and Tool Support," Conceptual Modeling – ER 2005, 2005, pp. 353-368.
- [19] N. Russell, Arthur, W. van der Aalst, e N. Mulyar, Workflow Control-Flow Patterns: A Revised View, 2006.
- [20] S.A. Macskassy, A. Banerjee, B.D. Davison, e H. Hirsh, "Human Performance on Clustering Web Pages: A Preliminary Study," KDD, 1998, pp. 264-268.
- [21] T. Berners-Lee, J. Hendler, e O. Lassila, "The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," 2001.
- [22] T.R. Gruber, "A translation approach to portable ontology specifications," Knowl. Acquis., vol. 5, 1993, pp. 199
- [23] Gomez-Perez, A., Knowledge sharing and reuse, 1998. In J. Liebowitz, ed., The Handbook of Applied Expert Systems, CRC Press.
- [24] Lucene, "<http://lucene.apache.org/>" Accessed in november 12th 2009.