

Classificação automática via ontologias: um estudo preliminar sobre raciocínio humano e lógica descritiva

Eduardo Ribeiro Felipe¹

Maurício Barcellos Almeida¹

¹Programa de Pós-Graduação em Gestão & Organização do Conhecimento
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

erfelipe@ufmg.br

mba@eci.ufmg.br

Abstract. *This paper presents an ongoing PhD project, which will be submitted to the examination board in February 2020, whose research question involves automatic classification via ontologies and human reasoning in description logics. The present research aims to investigate the classification in reasoners and compare it, from an empirical perspective, to results of human classification still very common in Information Science. At this moment, we extract some fragments from the literature that lists common problems people confront in modeling with description logics, and the methodology we will use to test the problems and suggest improvements.*

Resumo. *Esse artigo é resultado parcial de tese em andamento, com previsão de defesa em fevereiro de 2020, cujo problema de pesquisa envolve a classificação automática via ontologias. A presente pesquisa visa investigar a classificação em motores de inferência e compará-la de forma empírica com resultados de classificação manual, processo ainda comum em Ciência da Informação. Nesse momento, extraímos fragmentos da literatura que listam problemas comuns que as pessoas enfrentam ao modelar com lógica descritiva, além de apresentar a metodologia que deve ser aplicada para testar tais problemas e sugerir melhorias.*

1. Introdução

As ontologias são utilizadas na representação do conhecimento e foram popularizadas pela “Web Semântica”. Naquele contexto, computadores deveriam manipular conteúdos de forma semântica e realizar inferências [Berners-Lee et al., 2006]. A tecnologia buscava sistemas computacionais capazes de extrair conexões de diversas fontes de dados, estruturadas ou não, permitindo a produção de conteúdo não estático.

As lógicas descritivas (LDs) têm sido as linguagens mais usadas para a construção de ontologias. O desenvolvimento dos padrões OWL (Web Ontology Language) impulsionou ainda mais esse uso em função de razoável balanço que oferecem entre capacidade de expressão e capacidade de computação. No entanto, LDs são difíceis de compreender e trabalhar, especialmente, para usuários de ontologia que não são cientistas da computação e que não possuem treinamento em lógica. A sintaxe OWL Manchester foi desenvolvida para tornar as lógicas mais acessíveis ao adotar palavras-chave em inglês no lugar de símbolos lógicos. Ainda assim, as LDs continuam a apresentar dificuldades, mesmo quando representados na sintaxe de Manchester. A literatura relata diversas investigações sobre quais características das LDs causam dificuldades, como por exemplo, o contexto de entender como uma implicação decorre de uma justificação (o subconjunto mínimo da ontologia suficiente para tal vinculação se manter). No entanto, tem havido pouca pesquisa para relacionar essas dificuldades a como as pessoas naturalmente raciocinam e usam a linguagem, para fins de modelagem.

O presente trabalho se insere nesse contexto e é parte de uma pesquisa de doutorado em andamento, com previsão de defesa para fevereiro de 2020, cujo objetivo geral é entender as dificuldades que as pessoas experimentam ao usar LDs para criar, editar e avaliar ontologias, de forma a mitigar tais dificuldades. Os objetivos específicos são: a) construção de um software para interface com os motores de busca e inferência automática, b) criação de questionários e questões para comparação, c) submissão da mesma ontologia a estudantes de pós-graduação que trabalham com ontologias, d) comparação de resultados na classificação automatizada versus manual.

2. Classificação

Classificar remete a identificar e discernir características que tornem a entidade pertencente a determinado agrupamento. Este processo é feito de forma a permitir o reconhecimento por meio de características essenciais, que possibilitem a criação de um modelo descritivo das entidades ou conjunto delas.

Assim, o processo de classificação resulta em grupos conceituais genericamente denominados “classes”. Com as classes é possível organizar a informação de forma hierárquica e permitir que a especificidade de suas características seja evidenciada em subclasses. O ato de classificar, na CI, foi considerado por muitos anos um ato de “*arrumação sistemática*” [Barbosa, 1969], ou seja, realizar “um processo mental pelo qual coisas, são reunidas segundo as semelhanças ou diferenças que apresentam. Esta definição corrobora com Vickery (1975) e Simões (2011) que destacam o agrupamento por identificação de características comuns e sua separação por diferença.

No contexto da Ciência da Informação, o processo de classificação está diretamente ligado a organização de um acervo bibliográfico e sua recuperação. Com o advento da digitalização, o acervo é hoje digital, o que demanda o aprimoramento de recursos existentes, além do desenvolvimento de novas técnicas que permitam classificar conteúdos eletrônicos de forma automatizada como tem demonstrado a proposta do aprendizado de máquina.

3. Mecanismos de inferência

Na concepção da World Wide Web, o nível de representação permitiria que a informação fosse interpretada por agentes computacionais (algoritmos) e possibilitasse melhorias na recuperação da informação.

A tecnologia RDF foi uma das primeiras respostas na busca por expressividade para a Web Semântica. Sua estrutura compreende em três componentes básicos: *subject* (assunto), *predicate* (predicado) *and object* (objeto), formando o que se convencionou chamar de tripla [Powers, 2003]. Uma tripla é um fato, uma declaração (*assertion*), e pode ser conectada a outras triplas formando uma corrente de conexões, ampliando um modelo de representação do conhecimento.

O RDF Schema (RDFS), que consiste de *um vocabulário de modelagem de dados para dados em RDF*” (W3C), fornece os recursos necessários para descrever objetos e propriedades de um esquema de um domínio específico [Powers, 2003]. Em seguida, a fim de permitir maior expressividade e suporte à inferência por *reasoners*, foi criada a *Web Ontology Language* (OWL). OWL é uma lógica descritiva, subconjunto da Lógica de Primeira Ordem. Nem tudo que é escrito em OWL é necessariamente uma ontologia, e nem toda ontologia é representada em OWL [Rector, Sottara, 2014].

Neste contexto, os *reasoners* realizam inferências sobre a estrutura da ontologia criada a fim de adicionar conhecimento novo à mesma [Motik et al., 2009].

Rector et al. (2004) relacionaram os principais problemas enfrentados por alunos ao se utilizar as LDs para modelagem:

1. Tendência em assumir que “only” (universal) apenas implica “some” (existencial);
2. Confusão entre “and” e “or”;
3. A combinação de 1 e 2, com o uso de “and” entre a criação de classes disjuntas;
4. Confusão entre “P some (not X)” e “not (P some X)”.

4. Metodologia

A metodologia de pesquisa envolve os seguintes passos: i) construção de um aplicativo para processamento de inferências acessível a usuários não conhecedores de lógica; ii) identificação de padrões de modelagem em LD e raciocínio associado; iii) aplicação de questionários; iv) estudo dos resultados e propostas para mitigar os problemas encontrados.

4.1 - Construção de software com interface gráfica, de alto nível para permitir a execução do *reasoner*

Embora o *reasoner* possa funcionar em um editor de OWL como, por exemplo, o Protégé, este trabalho pretende possibilitar seu funcionamento utilizando uma interface web, para permitir o uso por não especialistas em ontologias, realidade comum na CI.

Nesta etapa o algoritmo de *reasoner* deve realizar as inferências, gerando como resultado uma ontologia acrescida de novos termos e/ou declarações. Uma camada voltada para visualização também deve ser desenvolvida ou adequada ao projeto, permitindo a navegação pelo usuário em classes e relações. Neste momento da pesquisa, foi identificada a ferramenta WebVOWL [Lohmann, 2014] que apresenta a ontologia em formato gráfico utilizando uma conversão do formato OWL em JSON.

4.2 - Identificação de padrões de modelagem em LD e raciocínio associado

Para identificar as dificuldades das pessoas em lidar com LDs, valeu-se de pesquisas já realizadas e disponíveis na literatura. No presente momento da pesquisa, adota-se um extrato da análise de padrões de conteúdo proposta por Khan e Blomqvist (2010). Os autores identificaram padrões de conteúdo mais comuns. A tabela 1 inclui as características encontradas nesses padrões (exceto de propriedades de dados e restrições de cardinalidade).

Tabela 1 – Características das LDs mais comuns

	Caraterística da linguagem	<i>Manchester OWL Syntax</i>
Características de Classes	subsunção	<i>SubClassOf</i>
	equivalência de classes	<i>EquivalentTo</i>
	disjunção classes	<i>DisjointWith</i>
	asserção de classes	<i>Type</i>
	conjunção	<i>and</i>
	disjunção	<i>or</i>
Características de Propriedades	complemento	<i>not</i>
	faixa	<i>Range</i>
	domínio	<i>Domain</i>
	hieraquia	<i>SubPropertyOf</i>
	propriedades inversas	<i>InverseOf</i>
	propriedades transitivas	<i>Characteristics: Transitive</i>
	propriedades funcionais	<i>Characteristics: Functional</i>
Restrições	Propriedades simétricas	<i>Characteristics: Symmetric</i>
	Restrição Universal	<i>some</i>
	Restrição Existencial	<i>only</i>

Adaptado de Khan and Blomqvist (2010).

Tabela 2 – Padrões para componentes

<i>Class Object SubClassOf has</i>	<i>component only Object</i>
	<i>SubClassOf is_component_of only Object</i>
<i>Property has_part</i>	<i>Characteristics Transitive</i>
<i>Property is_part_of</i>	<i>Characteristics Transitive</i>
<i>InverseOf has_part</i>	
<i>Property has_component</i>	<i>SubPropertyOf has_part</i>
<i>Property is_component_of</i>	<i>SubPropertyOf is_part_of</i>
	<i>InverseOf has_comp</i>

Tabela 3 – Questões para raciocínio (padrão componentes)

<i>A has component B</i>
<i>B has component C</i>
$\Rightarrow A \text{ has part } C$

Tabela 4 – Passos para o raciocínio (padrão componentes)

Passo número:	Raciocínio
1	A tem-component B; tem-componente SubPropertyOf has_part; $\Rightarrow A \text{ has_part } B$
2	B tem_componente C; tem-componente SubPropertyOf has_part $\Rightarrow B \text{ has_part } C$
3	has_part Characteristics Transitive; A has_part B [1]; B has_part C [2] $\Rightarrow A \text{ has_part } C$

4.3 Aplicação de questionários

Os questionários estarão no próprio site junto ao mecanismo de inferência, e devem conter pelo menos os seguintes dados:

- Tempo de resposta;
- Conhecimento de lógica;
- Número de passos de raciocínio;
- Validade e precisão;
- Dificuldades registradas;
- Comparação.

Nesta etapa uma interface gráfica deve permitir que o usuário possa discernir a ontologia original, a fim de permitir comparações, análises e declarações. Esta interface poderá utilizar outros softwares como, por exemplo, WebVOWL ou utilizar estruturas hierárquicas como TreeView, viabilizando alternativas a uma representação puramente textual.

4.4 Análise dos resultados

Uma vez analisados os resultados obtidos nos passos anteriores por meio de um procedimento de comparação entre os processos manuais e automatizados, passa-se a uma avaliação do que pode ser feito no sentido de mitigar as dificuldades encontradas e auxiliar usuários leigos na modelagem e consultas.

5. Considerações finais

Neste momento da pesquisa, o software para submissão da ontologia em OWL para o *reasoner* encontra-se em estágio funcional de desenvolvimento, apresentando como resultado um novo arquivo OWL contendo as definições e inferências geradas pelo algoritmo de *reasoner*.

O software já aborda as seguintes funcionalidades e características: a) permite a manipulação de uma ontologia em OWL ou RDFS e exibe sua visualização textual; b) realiza a validação da ontologia a fim de verificar o funcionamento do algoritmo do *reasoner*; c) submete a ontologia ao *reasoner* e grava um novo arquivo contendo uma nova ontologia, adicionada de suas inferências.

Espera-se com este trabalho contribuir com o aprimoramento e disseminação das ontologias na CI, expandindo a utilização de *reasoners* para aplicações no âmbito da classificação automática e para o entendimento das dificuldades experimentadas pelas pessoas ao lidar com a modelagem em lógica descritiva.

Referências

- BARBOSA, A. P. Teoria e prática dos sistemas de classificação bibliográfica. 1969.
- BERNERS-LEE, T.; HANDLER, J.; LASSILA, O. **The Semantic Web**. Disponível em: <<http://link-galegroup.ez27.periodicos.capes.gov.br/apps/doc/A147823458/AONE?sid=googlescholar>>. Acesso em: 15 abr. 2018.
- KHAN, M. T., BLOMQUIST, E. (2010). Ontology design pattern detection-initial method and usage scenarios. In *SEMAPRO 2010, The Fourth International Conference on Advances in Semantic Processing* (pp. 19–24).
- LOHMANN, S. et al. **WebVOWL: Web-based Visualization of Ontologies**. Knowledge Engineering and Knowledge Management. Springer, Cham, 24 nov. 2014 Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-17966-7_21>. Acesso em: 9 abr. 2018
- MOTIK, B., PATEL-SCHNEIDER, P.F., Parsia, B., 2009. OWL 2 Web Ontology Language Profiles. Available from: <<http://www.w3.org/TR/owl2-profiles/>>.
- POWERS, S. **Practical RDF**. Beijing; Sebastopol: O'Reilly, 2003.
- RECTOR, A., DRUMMOND, N., HORRIDGE, M., ROGERS, J., KNUBLAUCH, H., Stevens, R., ... WROE, C. (2004). OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *Engineering Knowledge in the Age of the Semantic Web* (pp.63–81). Springer.
- RECTOR, A.; SOTTARA, D. Formal Representations and Semantic Web Technologies. In: **Clinical Decision Support**. [s.l.] Elsevier, 2014. p. 551–598.
- SIMÕES, M. DA G. **Classificações bibliográficas: percurso de uma teoria**. Coimbra: Almedina, 2011.

SINGH, S.; KARWAYUN, R. **A Comparative Study of Inference Engines**. IEEE, 2010 Disponível em: <<http://ieeexplore.ieee.org/document/5501457/>>. Acesso em: 3 abr. 2018.

VICKERY, B. C. Ontologies. **Journal of Information Science**, v. 23, n. 4, p. 277–286, ago. 1997.