CARACTERIZAÇÃO DE DOCUMENTOS RDF

Lucas Pires Cobucci [1], Rebeca Schroeder Freitas [2]

[1] Acadêmico(a) do Curso de BCC bolsista PROIP/UDESC
[2] Orientadora, DCC rebeca.schroeder@udesc.br

Palavras-chave: RDF, Modelos Conceituais, Gerenciamento de Dados.

RDF (*Resource Description Framework*) é um modelo padrão para intercâmbio de dados na Web. Um documento RDF é formado por triplas, cada qual contendo 3 elementos: um sujeito, um predicado e um objeto. Esse modelo vem sendo utilizado para a descrição conceitual ou de modelagem de informações de recursos Web e no gerenciamento de dados de aplicações. Para se determinar a melhor forma de gerenciamento destes dados em um banco de dados, é necessário conhecer o esquema de um documento RDF. Apesar do RDF ser um modelo livre de esquema, é possível extrair informações sobre a estrutura dos dados. Estudos estão sendo conduzidos com intuito de apresentar metodologias para o armazenamento de documentos RDF e em geral, a utilização do modelo relacional vem sendo a mais usual, devido ao fato de que as triplas são mapeadas diretamente para uma tabela que contém estes 3 campos. Uma outra maneira de armazenamento é analisar a estrutura dos dados do documento em questão, de forma que seja possível criar um esquema relacional capaz de organizar os atributos de uma mesma classe de dados. Todavia, a compreensão da estrutura é representada no modelo relacional, o que limita a representação de dados em outros modelos de banco de dados. Com base nisso, este trabalho propõe a extração da estrutura de um documento RDF através de sua representação em um modelo conceitual e sua caracterização através de métricas que medem a homogeneidade de sua estrutura. O objetivo da extração de um modelo conceitual é favorecer a abstração dos dados, e habilitar futuramente o mapeamento dos dados para qualquer modelo lógico de banco de dados. Além disto, a compreensão da homogeneidade da estrutura dos dados poderá apoiar a escolha do modelo lógico mais adequado para implementação.

Este trabalho compreendeu as seguintes atividades: primeiramente foi realizada uma revisão da literatura para embasar o trabalho. Em seguida, deu-se andamento ao trabalho iniciado em [2], para que fosse possível a extração de modelos conceituais e suas métricas. Na atividade subsequente, foi feita uma seleção de documentos RDF que seriam avaliados pelo extrator desenvolvido. Nessa etapa foi necessário fazer o *download* de conjuntos de dados (*datasets*) da Internet. Dentre todos os selecionados, 8 foram *datasets* reais e 4 foram criados a partir de *benchmarks*. Posteriormente foi realizada a análise em relação a homogeneidade da estrutura dos documentos RDF, que basicamente envolveu avaliar qual a porcentagem das instâncias de uma entidade presente no conjunto de dados apresenta os atributos da entidade. Este mesmo tipo de análise foi realizada para os relacionamentos do esquema. Em ambos os casos, a avaliação foi realizada através da extração de métricas que determinaram as cardinalidades reais de atributos e relacionamentos do esquema extraído em relação a um documento RDF. Para fins de validação das métricas implementadas, foi realizada a comparação com métricas de um trabalho fortemente relacionado [1]. Neste caso, foi utilizado a métrica *coherence* que mede o nível de estruturação de um documento RDF, cujos valores do cálculo variam de 0 até 1, indicando que quanto mais perto de 1 mais estruturado o documento será.

Um dos problemas encontrados neste trabalho foi a obtenção de documentos RDF, tendo em vista que muitos não possuíam uma quantidade significativa de relacionamentos para avaliação ou não estavam disponíveis para *download*. Dos 20 documentos selecionados para extração, apenas 12 foram analisados. Esquemas conceituais foram extraídos para cada documento, assim como representado na Figura 1, onde triplas RDF estão sendo modeladas para um esquema conceitual (esquema ER). Um exemplo seriam as triplas *ProductFeature2-type-ProductFeature* e *ProductFeature2-label-Classic* que estão sendo modeladas no esquema ER de forma que *ProductFeature2* é uma instância da entidade *ProductFeature* e

esta apresenta o atributo *label*, que é uma característica (atributo) de *ProductFeature2* e tem valor (atribuição) denominado "Classic". O mesmo vale para o relacionamento entre as entidades *ProductFeature* e *Product*, que é denominada *feature*. Neste caso, como *Productfeature2* é uma instância de *ProductFeature* e *Product1* é uma instância de *Product*, entende-se que existe um relacionamento entre as entidades *ProductFeature* e *Product*. As métricas extraídas como o mínimo e máximo da cardinalidade de atributos e relacionamentos indicaram a homogeneidade da estrutura dos documentos RDF. A Tabela 1 apresenta os resultados para as cardinalidades mínimas para todos os *datasets* avaliados. Por exemplo, no *dataset Berlin 50,* dos 43 atributos deste esquema, 23% são opcionais indicado pela cardinalidade mínima ser inferior a 1 (*min_card < 1*). A mesma análise foi feita para os relacionamentos, em que se avaliou a cardinalidade mínima de todas as entidades participantes nos relacionamentos. Em geral, os resultados demonstraram que os documentos criados por *benchmarks* (Berlin) apresentam estruturas mais homogêneas com porcentagem do *min_card* inferiores, se comparado aos *datasets* reais. Algumas exceções ocorreram para os *datasets* reais GDPR e R4R, em que se observou valores reduzidos para *min_card.*

	Atributos		Relacionamentos		
Datasets	#	% min_card < 1	#	% min_card < 1	Coherence
Berlin 50	43	23%	16	31%	0.965
Berlin 100	43	23%	16	31%	0.958
Berlin 500	43	23%	16	37%	0.953
Berlin 1000	43	23%	16	37%	0.952
Court	26	57%	44	90%	0.808
DSA	9	77%	44	91%	0.176
GDPR	49	6%	214	44%	0.845
PPO	21	57%	68	63%	0.435
R4R	18	22%	24	91%	0.794
STW	52	38%	86	78%	0.642

Tabela 1 – Métricas dos *datasets* avaliados

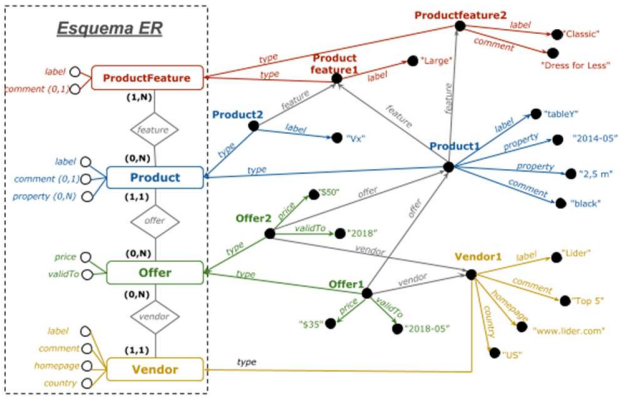

Figura 1 – Exemplo de um mapeamento de RDF para esquema conceitual (esquema ER)

Ao comparar as métricas produzidas por este trabalho com a métrica *coherence* proposta em [1], é possível constatar pela Tabela 1 que as métricas são compatíveis. Ou seja, *datasets* com *min_card* com valores mais baixos apresentam valores para o *coherence* mais próximos a 1, indicando *datasets* mais homogêneos. Entretanto, o diferencial do trabalho desenvolvido é permitir representar tais métricas para cada atributo e relacionamento de um esquema conceitual, enquanto em [1] a medida é dada para o esquema como um todo. Alguns valores de cardinalidades podem ser observados no esquema ER presente na Figura 1.

As conclusões deste trabalho foram que a partir dos resultados desta pesquisa é possível extrair esquemas conceituais de qualquer documento que esteja no formato RDF, bem como calcular métricas que determinam o grau de estruturação de seus componentes. Através dos experimentos realizados foi possível caracterizar um conjunto de documentos RDF provenientes de *benchmarks* e de aplicações reais. Os resultados produzidos por este tipo de análise são capazes de apoiar a decisão pelo formato mais apropriado para o gerenciamento destes dados, de forma que seja possível mapear as características estruturais dos documentos para um modelo lógico de banco de dados compatível.

[1] Duan, S., Kementsietsidis, A., Srinivas, K., and Udrea, O. (2011). Apples and oranges:A comparison of rdf benchmarks and real rdf datasets. InProceedings of the 2011ACM SIGMOD International Conference on Management of Data, SIGMOD '11, pages 145–156.
[2] MAIA, Alisson. Uma abordagem para extração de esquemas RDF. 2015. Trabalho de conclusão de curso (Bacharelado em Ciência da Computação) - Universidade do Estado de Santa Catarina, Joinville, 2015.