Impacto da Evolução de Dados Interconectados Abertos

Julio K. R. Matsoui, Julio C. Dos Reis.

Resumo

A descrição de dados abertos em formato padrão sobre entidades na Web tem tido uma grande adoção nos últimos anos. O principal formato explorado na definição de elementos de dados é o *Resource Description Framework* (RDF). O número de repositórios de dados nesse formato provendo dados interconectados entre diferentes fontes tem se tornado numeroso. O grande volume de dados exige cada vez mais métodos e ferramentas automáticas para efetuar análises e correções sobre os dados. Em particular, dados interconectados no contexto da Web Semântica tendem a ser dinâmicos. Novas versões de conceitos, suas relações e instâncias são redefinidos ao longo do tempo e de certa maneira alteram o significado e propriedades das entidades. Logo, tornam-se necessárias técnicas para se entender a evolução de dados interconectados e como tal evolução influencia em interconexões estabelecidas entre bases distintas.

Palavras-chave:

LOD, Link Evolution, Change Operations.

Introdução

A característica fundamental da Web reside no fato da existência de links que apontam de um documento para outro. No contexto da Web Semântica, esses links expressam interconexão entre elementos de dados estruturados e cuja semântica é explicitamente definida. Assumindo que as propriedades das entidades descritas em repositórios de dados expressos em RDF² podem sofrer modificações, elas podem impactar em conexões (links) já estabelecidas. O grande volume de dados atualmente nos repositórios de dados da Web Semântica impede que correções manuais sejam efetuadas em situações onde modificações invalidam conexões previamente definidas.

Resultados e Discussão

Para o estudo da evolução dos dados abertos interligados, utilizou-se o dataset do Agrovoc¹. Neste dataset foi observado as mudanças simples e complexas que ocorreram entre as versões de abril de 2018 (D^j) e abril de 2019 (D^{j+1}). O total de triplas em D^j é de 4.254.655 e em D^{j+1} temos um total de 4.540.205 triplas. O Diff, (D^j , D^{j+1}), detectou 202 remoções, 6.990 adições e 528.275 modificações. Logo, percebe-se que as modificações tem um papel mais importante nas mudanças ocorridas de uma versão para outra. Para isso, aprofunda-se o estudos das modificações separando-as em quatro grupos, a saber: M:1 - modificação da tripla interna acarretando uma adição de um link externo; M:2 - modificação da tripla interna acarretando em uma remoção de um link externo; M:3 - modificação da tripla interna acarretando em uma modificação no link externo; M:4 - modificação da tripla interna sem modificar o link externo. Para fazer o experimento, primeiramente deve-se calcular o Diff e posteriormente rodar o análises sobre este. Assim, obteve-se as seguinte tabela com os resultados.

Tabela 1. Resultado da análise da modificação.

Tipo	Total de ocorrências
M:1	0
M:2	245
M:3	21.899
M:4	474.253

Da Tabela 1, nota-se que a maioria das ocorrências encontra-se em M:3 e M:4, representando 99.96% das modificações. Diante disso, percebe-se que o dataset tende a preservar seus links.

Conclusão

Este trabalho visou estudar os efeitos das mudanças sobre os links externos. A motivação deste estudo é seu caráter inovativo frente ao parco material publicado sobre o assunto. Entretanto, ele tem um papel importante na manutenção dos dados interconectados abertos na internet. Diante dos experimentos propostos neste trabalho, chegamos em alguns resultados interessantes, como o papel fundamental da modificação, mudança complexa, na evolução dos dados; A maioria dos dados no dataset tende a se manter estável, ou seja, o tamanho do Diff é muito pequeno em relação ao tamanho total do dataset.

Agradecimentos

Gostaria de agradecer às seguintes pessoas: Meus amigos de república que sempre me mantiveram motivado na realização deste trabalho; meu orientador Julio Cesar dos Reis; e, por fim, à FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo, que financiou a pesquisa que deu origem ao artigo científico, processo nº 2018/05357-8.

¹ RDF, W3C <Disponível em: <https://www.w3.org/RDF/>>, último acesso no dia 01/08/2019..

² Dataset, AGROVOC <Disponível em: <http://aims.fao.org/agrovoc/releases>>, último acesso no dia 17/03/2019