3 Trabalhos Relacionados

Este capítulo apresenta trabalhos relacionados que se preocupam com a criação das relações entre os dados publicados na Web.

3.1 Silk Framework

Volz, Bizer, Gaedke, & Kobilarov (2009) apresentam um framework para a descoberta de relações entre entidades das diferentes fontes publicadas na Web de Dados. Este framework apoia os editores na definição de ligações RDF entre suas fontes de dados e outras disponíveis na Web. A partir de uma linguagem declarativa, é possível especificar que tipos de ligações RDF deverão ser descobertas entre as fontes de dados mapeadas, bem como as condições e métricas de similaridade que os dados devem cumprir para serem interligados. Silk acessa fontes de dados através de consultas do tipo SPARQL Protocol and RDF Query Language (SPARQL) e pode assim ser utilizado para identificar ligações entre fontes de dados locais e remotas.

É uma ferramenta de tempo otimizado. Em vez de usar as características dos espaços métricos, utiliza um índice de correspondência prévia para chegar a uma complexidade de tempo quase linear. Para isso, antes de executar quaisquer comparações, obtém a lista de entidades da fonte de origem e da fonte a ser consultada, e a armazena no disco para reutilização em uma nova execução (cache). Essa lista é utilizada para montar o índice de correspondência. Durante a comparação entre as duas fontes, procura no índice as entidades candidatas para cada entidade de origem, limitando a quantidade de comparações detalhadas. Esse “prematching” é opcional, mas recomendado, pois reduz drasticamente o tempo de execução e carga de rede.

Durante a execução de cada comparação detalhada de pares de entidades, as métricas de agregação definidas pelo usuário são avaliadas e os resultados da consulta são armazenados na memória.

Se uma dessas métricas de agregação retornar um valor acima dos limiares especificados, a ligação candidata é guardada na memória.

Depois de completar todas as comparações definidas, um limite pode ser aplicado para determinar o número máximo de ligações obtidas a partir de uma única entidade. Apenas uma quantidade especificada de ligações mais cotadas será mantida, descartando as menos valorizadas.

3.2 Limes

LIMES é um framework para descoberta de ligações na Web de Dados, considerado uma abordagem eficiente para a descoberta de ligações em larga escala. Este framework aborda o problema da escalabilidade utilizando uma solução eficiente, com técnicas que calculam estimativas pessimistas de semelhanças entre instâncias (Ngonga Ngomo & Auer, 2011).

A partir de condições configuradas pelo usuário, e com base nas estimativas calculadas, o framework filtra um grande número de pares de instâncias que não se encaixam nas condições definidas. Os pares restantes têm suas semelhanças reais calculadas, e as instâncias finais são devolvidas. Com isso, utiliza uma quantidade menor de comparações, o que caracteriza a eficiência dessa solução.

O framework também pode ser baixado como ferramenta independente para descobrir ligações localmente (Ngonga Ngomo, 2013).

3.3 Sindice.com

Sindice é um índice semântico composto por serviços de busca sobre entidades obtidas na Web Semântica, que ajuda a encontrar, entender e integrar Web sites.

Este índice permite que aplicações obtenham automaticamente fontes de dados que possuem informações sobre uma determinada entidade. Também permite a obtenção de entidades através de propriedades do tipo inverse-functional e de pesquisas de texto completo, além de indexar SPARQL Endpoints (Tummarello, Oren, & Delbru, 2007).

Tummarello (2011) define como uma forma consolidada e personalizável de acessar os dados disponíveis na Web. Com metadados de mais de 260 milhões de páginas em seu índice, Sindice.com oferece o maior repositório mundial, continuamente atualizado, do conteúdo da Web de Dados. Suas principais características incluem: um crawler focado que permite indexação rápida de dados; e índices de dados proprietários que são atualizados em tempo real.

3.4 Reconciliação de entidades entre fontes da Web de Dados

Maali, Cyganiak, & Peristeras (2011) comparam abordagens de consulta a entidades através do uso da linguagem SPARQL; de técnicas de ontology matching; e de buscas de texto sobre dados RDF. O termo reconciliação é utilizado para definir a comparação de duas entidades utilizando uma abordagem de similaridade.

O trabalho compara as abordagens abaixo, que fornecem uma API bem definida para acesso remoto, assim como implementações que permitem suas aplicações em diferentes fontes de dados RDF.

- a) **SPARQL**: a linguagem de consultas RDF recomendada pelo World Wide Web Consortium (W3C) (Prud'hommeaux & Seaborne, 2008);
- b) **SPARQL with full-text search**: a linguagem SPARQL com extensões adicionais de busca em texto completo;
- c) **Silk Server**: baseado no Silk Framework, fornece uma interface RESTful (Wilde & Hausenblas, 2009) para conectar dois conjuntos de dados;
- d) **Sindice**: mecanismo de busca semântica na Web de Dados.

Os resultados das comparações favorecem o Silk Server, por ser especialista em interligações, baseada em métricas de similaridade, mas dependente da complexidade dos dados referenciados.