Pontifícia Universidade Católica
do Rio de Janeiro

José Eduardo Talavera Herrera

Arquitetura para Recomendação de Fontes de Dados RDF

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do grau de Mestre em Informática.

Orientadora: Profª. Karin Breitman

Rio de Janeiro
Outubro de 2012

Pontifícia Universidade Católica
do Rio de Janeiro

José Eduardo Talavera Herrera

Arquitetura para Recomendação de Fontes de Dados RDF

Dissertação apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do grau de Mestre em Informática. Aprovada pela Comissão Examinadora abaixo assinada.

Profª. Karin Breitman
Orientadora
Departamento de Informática — PUC–Rio

Prof. Marco Antonio Casanova
Departamento de Informática — PUC-Rio

Prof. Luiz André P. Paes Leme
UFF-Rio

Prof. Antonio Luz Furtado
Departamento de Informática — PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro Técnico Científico — PUC–Rio

Rio de Janeiro, 01 de Outubro de 2012

José Eduardo Talavera Herrera

Graduou-se em Engenharia de Sistemas Universidad Nacional de San Agustín (UNSA), Arequipa-Perú em 2003. Trabalha como Engenheiro de Aplicações no Banco de Credito do Perú. Ingressou no programa de mestrado do Departamento de Informática em 2010. Atualmente suas áreas de pesquisa são Web Semântica, *Linked Data* e *Cloud Computing*.

Ficha Catalográfica

Herrera, José Eduardo

Arquitetura para Recomendação de Fontes de Dados RDF / José Eduardo Talavera Herrera; orientadora: Karin Breitman. — 2012.

112 f. : il. (color); 30 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2012.

Inclui bibliografia.

1. Informática – Teses.
2. Dados Conectados. 3. Computação na nuvem. 4. Recuperação de informação. 5. Descoberta de Links. 6. Similaridade. I. Breitman , Karin. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Agradeço a Deus por ter me dado força, saúde e por ter estado ao meu lado sempre me guiando. Agradeço aos meus pais, Soledad e Eduardo, pelo amor, dedicação, amizade e apoio e à minha irmã Diana pelo amor, amizade, incentivo, ajuda e apoio em todos momentos. Agradeço a Karin, minha orientadora, por ter me recebido no seu grupo de pesquisa, pelas oportunidades e pela orientação. Agradeço aos meus amigos Renato, Ricardo, Marcelo, Chrystiano e Everton pela amizade, apoio e por tornar os momentos sempre muito divertidos. Agradeço a Michele por ter estado do meu lado e por sempre acreditar em mim, você é e será muito especial para mim. Agradeço aos meus amigos Livia, Thiago e Alessandro pela ajuda na correção deste trabalho, muito obrigado. Agradeço a Amparito, Ximena, Mayra, Edward, Ronald e Adriano pelo incentivo, ajuda e amizade. Enfim, agradeço à todos os meus amgios pois foram peças fundamentais para que eu pudesse chegar até aqui. E Agradeço à CAPES que financiou minha bolsa de mestrado.

Resumo

Herrera, José Eduardo; Breitman , Karin. **Arquitetura para Recomendação de Fontes de Dados RDF**. Rio de Janeiro, 2012. 112p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Dentro do processo de publicação de dados na Web recomenda-se interligar os dados entre diferentes fontes, através de recursos similares que descrevam um domínio em comum. No entanto, com o crescimento do número dos conjuntos de dados publicados na Web de Dados, as tarefas de descoberta e seleção de dados tornam-se cada vez mais complexas. Além disso, a natureza distribuída e interconectada dos dados, fazem com que a sua análise e entendimento sejam muito demorados. Neste sentido, este trabalho visa oferecer uma arquitetura Web para a identificação de fontes de dados em RDF, com o objetivo de prover melhorias nos processos de publicação, interconexão, e exploração de dados na *Linked Open Data*. Para tal, nossa abordagem utiliza o modelo de MapReduce sobre o paradigma de computação nas nuvens. Assim, podemos efetuar buscas paralelas por palavras-chave sobre um índice de dados semânticos existente na Web. Estas buscas permitem identificar fontes candidatas para ligar os dados. Por meio desta abordagem, foi possível integrar diferentes ferramentas da web semântica em um processo de busca para descobrir fontes de dados relevantes, e relacionar tópicos de interesse definidos pelo usuário. Para atingir nosso objetivo foi necessária a indexação e análise de texto para aperfeiçoar a busca de recursos na *Linked Open Data*. Para mostrar a eficácia de nossa abordagem foi desenvolvido um estudo de caso, utilizando um subconjunto de dados de uma fonte na *Linked Open Data*, através do seu serviço SPARQL *endpoint*. Os resultados do nosso trabalho revelam que a geração de estatísticas sobre os dados da fonte é, de fato, um grande diferencial no processo de busca. Estas estatísticas ajudam ao usuário no processo de escolha de indivíduos. Um processo especializado de extração de palavras-chave é aplicado para cada indivíduo com o objetivo de gerar diferentes buscas sobre o índice semântico. Mostramos a escalabilidade de nosso processo de recomendação de fontes RDF através de diferentes amostras de indivíduos.

Palavras–chave

Dados Conectados; Computação na nuvem; Recuperação de informação; Descoberta de Links; Similaridade.

Abstract

Herrera, José Eduardo; Breitman , Karin (Advisor). **An Architecture for RDF Data Sources Recommendation** . Rio de Janeiro, 2012. 112p. MSc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

In the Web publishing process of data it is recommended to link the data from different sources using similar resources that describe a domain in common. However, the growing number of published data sets on the Web have made the data discovery and data selection tasks become increasingly complex. Moreover, the distributed and interconnected nature of the data causes the understanding and analysis to become too prolonged. In this context, this work aims to provide a Web architecture for identifying RDF data sources with the goal of improving the publishing, interconnection, and data exploration processes within the Linked Open Data. Our approach utilizes the MapReduce computing model on top of the cloud computing paradigm. In this manner, we are able to make parallel keyword searches over existing semantic data indexes available on the web. This will allow to identify candidate sources to link the data. Through this approach, it was possible to integrate different semantic web tools and relevant data sources in a search process, and also to relate topics of interest defined by the user. In order to achieve our objectives it was necessary to index and analyze text to improve the search of resources in the Linked Open Data. To show the effectiveness of our approach we developed a case study using a subset of data from a source in the Linked Open Data through its SPARQL endpoint service. The results of our work reveal that the generation and usage of data source's statistics do make a great difference within the search process. These statistics help the user within the choosing individuals process. Furthermore, a specialized keyword extraction process is run for each individual in order to create different search processes using the semantic index. We show the scalability of our RDF recommendation process by sampling several individuals.

Keywords

Linked Data; Cloud computing; Retrival Information; Link Discovery; Similarity.

Sumário

Lista de figuras

Lista de tabelas