Pontifícia Universidade Católica
do Rio de Janeiro

Danilo Moret Rodrigues

Distributed RDF Graph Keyword Search

Dissertação de Mestrado

Dissertation presented to the Programa de Pós-Graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Karin Koogan Breitman

Rio de Janeiro

March 2013

Danilo Moret Rodrigues

Distributed RDF Graph Keyword Search

Dissertation presented to the Programa de Pós-Graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio, as partial fullfilment of the requirements for the degree of Mestre.

Prof. Karin Koogan Breitman
Advisor
Departamento de Informática – PUC-Rio

Prof. Marco Antônio Casanova
Departamento de Informática - PUC-Rio

Prof. José Viterbo Filho
UFF

Prof. Antônio Luz Furtado
Departamento de Informática - PUC-Rio

Prof. José Eugenio Leal
Coordinator of the Centro Técnico Científico – PUC-Rio

Rio de Janeiro, March 25th, 2013

Danilo Moret Rodrigues

Graduated in Telecommunications Engineering from Universidade Federal Fluminense - UFF in 2005.

Bibliographic data

Rodrigues, Danilo Moret

Distributed RDF Graph Keyword Search / Danilo Moret ; advisor: Karin Koogan Breitman. – 2013.

66 f. : il. ; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2013.

Inclui bibliografia

1. Informática – Dissertações. 2. Redes de Computadores e Sistemas Distribuídos. 3. Computação em Nuvem. 4. Linked Data. 5. Computação Distribuída. I. Breitman, Karin. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Matemática. III. Busca distribuída em grafo RDF por palavra-chave.

CDD: 004

Acknowledgements

To Graça, Karin and Renata.

Abstract

Moret, Danilo; Breitman, Karin. **Distributed RDF Graph Keyword Search**. Rio de Janeiro 2013. 66pp. MSc Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The goal of this dissertation is to improve RDF keyword search. We propose a scalable approach, based on a tensor representation that allows for distributed storage, and thus the use of parallel techniques to speed up the search over large linked data sets, in particular those published as Linked Data. An unprecedented amount of information is becoming available following the principles of Linked Data, forming what is called the Web of Data. This information, typically codified as RDF subject-predicate-object triples, is commonly abstracted as a graph which subjects and objects are nodes, and predicates are edges connecting them. As a consequence of the widespread adoption of search engines on the World Wide Web, users are familiar with keyword search. For RDF graphs, however, extracting a coherent subset of data graphs to enrich search results is a time consuming and expensive task, and it is expected to be executed on-the-fly at user prompt. The dissertation's goal is to handle this problem. A recent proposal has been made to index RDF graphs as a sparse matrix with the pre-computed information necessary for faster retrieval of sub-graphs, and the use of tensor-based queries over the sparse matrix. The tensor approach can leverage modern distributed computing techniques, e.g., non-relational database sharding and the MapReduce model. In this dissertation, we propose a design and explore the viability of the tensor-based approach to build a distributed datastore and speed up keyword search with a parallel approach.

Keywords

Linked Data, MapReduce, Cloud Computing, Keyword Search

Resumo

Moret, Danilo; Breitman, Karin. **Busca distribuída em grafo RDF por palavra-chave**. Rio de Janeiro 2013. 66pp. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

O objetivo desta dissertação é melhorar a busca por palavra-chave em formato RDF. Propomos uma abordagem escalável, baseada numa representação tensorial, que permite o armazenamento distribuído e, como consequência, o uso de técnicas de paralelismo para agilizar a busca sobre grandes bases de RDF, em particular, as publicadas como Linked Data. Um volume sem precedentes de informação está sendo disponibilizado seguindo os princípios de Linked Data, formando o que chamamos de Web of Data. Esta informação, tipicamente codificada como triplas RDF, costuma ser representada como um grafo, onde sujeitos e objetos são vértices, e predicados são arestas ligando os vértices. Em consequência da ampla adoção de mecanismos de busca na World Wide Web, usuários estão familiarizados com a busca por palavra-chave. No caso de grafos RDF, no entanto, a extração de uma partição coerente de grafos para enriquecer os resultados da busca é uma tarefa cara, demorada, e cuja expectativa do usuário é de que seja executada em tempo real. Este trabalho tem como objetivo o tratamento deste problema. Parte de uma solução proposta recentemente prega a indexação do grafo RDF como uma matriz esparsa, que contém um conjunto de informações pré-computadas para agilizar a extração de seções do grafo, e o uso de consultas baseadas em tensores sobre a matriz esparsa. Esta abordagem baseada em tensores permite que se tome vantagem de técnicas modernas de programação distribuída, e.g., a utilização de bases de dados não-relacionais fracionadas e o modelo de MapReduce. Nesta dissertação, propomos o desenho e exploramos a viabilidade da abordagem baseada em tensores, com o objetivo de construir um depósito de dados distribuído e agilizar a busca por palavras-chave com uma abordagem paralela.

Palavras-chave

Linked Data, MapReduce, Cloud Computing, Keyword Search

Contents

Figures

Algorithms

Tables