Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Fuzzy Sets and Systems ••• (••••) •••–•••

FUZZY sets and systems

www.elsevier.com/locate/fss

Handling scalable approximate queries over NoSQL graph databases: Cypherf and the Fuzzy4S framework

Arnaud Castelltort [a], Trevor Martin [b]

[a] *Montpellier, France*
[b] *Department of Engineering Mathematics, University of Bristol, UK*

Received 11 June 2016; received in revised form 30 July 2017; accepted 6 August 2017

Abstract

NoSQL databases are currently often considered for Big Data solutions as they offer efficient solutions for volume and velocity issues and can manage some of complex data (e.g., documents, graphs). However, fuzzy approaches are often not efficient on such frameworks. Thus this article introduces a novel approach to define and run approximate queries over NoSQL graph databases using Scala by proposing the Fuzzy4S framework and the Cypherf fuzzy declarative query language. NoSQL Graph databases are currently gaining more and more interest and are applied in many real world applications. The Fuzzy4S framework is defined with an open DSL (Domain Specific Language) allowing it to define scalable approximate queries at an abstract level. Cypherf is an extension of Cypher which runs over the Neo4J NoSQL graph databases. This work consists of a complete approach embedding the whole chain from end-user declarative query level to implementation issues within the database engine. We provide both the formal definitions for defining approximate graph NoSQL queries and the experimental results which demonstrate the interest and efficiency of our proposition.

Keywords: Approximate queries; Fuzzy logic; NoSQL graph databases

1. Introduction

Graphs are often a natural way to represent the world [1]. This is especially true for social networks, but also in biology, chemistry etc. In such areas, relationships between objects are at least as important as the objects themselves. For example, retrieving “friend” and “friend of friend” relationships is a key operation in social networks or in any other human relations. In [2], we have highlighted the powerness of the study of relations in order to retrieve fraud rings which are defined as *sophisticated chains of indirect links between fraudsters representing successive transactions (money, communications, etc.) from which rogue behaviors are detected.*

E-mail address: arnaud.castelltort@lirmm.fr (A. Castelltort).

http://dx.doi.org/10.1016/j.fss.2017.08.002
0165-0114/© 2017 Published by Elsevier B.V.

19 systems in ranking, November 2015

Rank			DBMS	Database Model	Score		
Nov 2015	Oct 2015	Nov 2014			Nov 2015	Oct 2015	Nov 2014
1.	1.	1.	Neo4j	Graph DBMS	34.04	+0.63	+9.39
2.	2.	2.	Titan	Graph DBMS	6.06	+0.54	+3.45
3.	3.	3.	OrientDB	Multi-model	5.50	+0.57	+3.48
4.	4.	↑ 7.	ArangoDB	Multi-model	1.61	+0.13	+1.30
5.	5.	↑ 6.	Giraph	Graph DBMS	0.92	-0.00	+0.45
6.	6.	↓ 5.	AllegroGraph	Multi-model	0.91	+0.01	+0.30
7.	7.	↑ 11.	Stardog	Multi-model	0.54	+0.02	+0.41
8.	↑ 9.	↑ 9.	Sqrrl	Multi-model	0.42	-0.01	+0.24
9.	↓ 8.	↓ 8.	InfiniteGraph	Graph DBMS	0.38	-0.05	+0.12
10.	10.	↓ 4.	Sparksee	Graph DBMS	0.29	-0.02	-0.59
11.	11.	↑ 15.	HyperGraphDB	Graph DBMS	0.25	-0.01	+0.22
12.	12.	12.	InfoGrid	Graph DBMS	0.21	+0.00	+0.09
13.	↑ 14.		VelocityGraph	Graph DBMS	0.19	+0.02	
14.	↑ 15.	↑ 16.	GlobalsDB	Multi-model	0.18	+0.02	+0.18
15.	↓ 13.	↓ 13.	FlockDB	Graph DBMS	0.16	-0.02	+0.04
16.	↑ 17.	↓ 10.	GraphDB	Multi-model	0.10	+0.06	-0.06
17.	↓ 16.	↓ 14.	GraphBase	Graph DBMS	0.05	-0.01	+0.02
18.	18.		Blazegraph	Multi-model	0.01	+0.00	
19.	19.	↓ 16.	Amisa Server	Multi-model	0.00	±0.00	±0.00

Fig. 1. Comparison of graph database management systems (db-engines.com).

Graphs have been extensively studied in the literature and have recently gained attention with the development of Semantic Web and ontologies. Many modelizations, tools and frameworks have been suggested to represent, manage and analyze graphs with XML, OWL etc.

At the same time, the volume and complexity of information concerning these objects and their relationships are growing dramatically, leading to huge databanks.

This has led to the development of the so-called NoSQL graph databases that embed both the performances of NoSQL databases and the representativity of graphs. Several engines exist (OrientDB, Neo4J, HyperGraphDB etc.) [3]. In this article, Neo4J is used since it is often considered the top graph database management system[1] (Fig. 1, [4][2]).

Such database management systems embed query processing that can be used at different levels, from the declarative level to the programmatic level within the engine. The Neo4J system offers the declarative Cypher language that allows the users to query the graph in a very intuitive manner. These queries include clauses over both nodes and relationships. Properties can be defined in the same style as NoSQL databases *i.e.* with $(key; value)$ lists. The value can be of any type, including collections.

For instance, it is possible to consider nodes representing people having some properties such as age $(key = age, value = 28)$ or hobbies $(key = hobby, value = \{tennis, reading, travel\})$.

Retrieving people having friends aged between 20 and 30 is then easily achieved using the available query languages.

However, in many cases, it is not possible or not relevant for users to define such queries in a “crisp” way. For instance it may be important to retrieve some people even if their age is 31 and not exactly between 20 and 30.

For this reason, we consider approximate queries. Such queries can for instance be expressed using the fuzzy set theory framework. In our proposition we introduce Fuzzy4S, standing for *Fuzzy for Scala*. Scala is a recent language mixing object-oriented and functional programming [5]. Fuzzy4S is used within NoSQL databases to define fuzzy queries by introducing the Cypherf framework (which stands for *Cypher fuzzy*).

The rest of this article is organized as follows: Section 2 presents the preliminary concepts underlying our work: graph databases and approximate queries. Section 3 introduces the extension of graph NoSQL queries with the Cypherf framework and discusses implementation issues. Cypherf relies on the Fuzzy4S framework which is detailed

[1] http://db-engines.com/en/ranking/graph+dbms.
[2] http://thoughtworks.fileburst.com/assets/technology-radar-may-2013.pdf.

Please cite this article in press as: A. Castelltort, T. Martin, Handling scalable approximate queries over NoSQL graph databases: Cypherf and the Fuzzy4S framework, Fuzzy Sets Syst. (2017), http://dx.doi.org/10.1016/j.fss.2017.08.002