Abschluss Arbeit

Neo4j und..

Asmaa Haja	Ilhem Bouzir	Josseph
348758	350778	330187

6. Juli 2015

Muss geaendert werden: TU Berlin – Fakultät IV
Institut für Wirtschaftsinformatik und Quantitative Methoden
Fachgebiet Informatik und Gesellschaft
Prof. Dr.-Ing. Frank Pallas
Max Ulbricht

Inhaltsverzeichnis

Abbildungsverzeichnis

Listings

1 Abstract

Dieser Bericht beschreibt die Erstellung eines knowledgegraphs mit Hilfe von Neo4j. Dieser Graph repräsentiert verschiedene Deutsche städte ,deren Sehenswürdigkeiten und deren Entfernungen voneinander. Das Projekt fand im Rahmen der Veranstaltung DBPRO an der TU Berlin statt.

2 Einleitung

Im Bereich Big Data fallen zum Beispiel bei Texten sehr große Daten-mengen an. Sinnvollerweise sollten solche Text Informationen auf einer übersichtlicherer und anschaulischerer Weise dargestellt werden können, so dass anhand einer bestimmten Repräsentation der Daten aus diesem Text gewonnenwerden. Diese erlaubt eine leichtere Suche nach bestimmten Informationen, ohne dass dabei wichtige Daten verloren gehen. Ziel des Projektes war es eine Knowledge Graph mit Hilfe der Graphendatenbank Neo4j zu erstellen basierend auf bestimmte rausgesuchte Daten.Mit der Graphenbasierte Abfragesprache sollen dann benötigte Informationen aus einer Wikipedia Informationen basierte Linked Open Data Version (DBpedia) extrahiert werden können und in einem bestimmten Format zu Neo4j übergen werden, so dass diese die Daten im strukturierten Graphen Form automatisch bringt. Das Projekt fand im Sommersemester 2015 an der TU Berlin im Rahmen der Veranstaltung DBPRO statt. Das Projektteam bestand aus Asmaa Haja, Ilhem Bouzir und Josseph, welche durch Dr. Holmer Hemsen betreut wurden.Das Projekt umfasste drei Meileinsteine. Zunächst fand im 1. Meilenstein eine Einarbeitung in das Thema Knowledge Graph, Neo4j und SPARQL statt. Als nächstes wurde dann konkrete Beispiele für Tourism in Deutschland mit SPARQL erstellt. Im 3. Meilenstein wurde aus der zuvor erstellten Abfragen mit Hilfe von Neo4j ein Knowledge Graph erstellt.Diese Arbeiten werden in diesem Bericht nach einer genauen Identifizierung der Aufgaben näher erläutert. Abschließend folgen ein Kapitel zur Evaluation und eine kurze Zusammenfassung der Projektergebnisse.

3 Aufgaben

Als nächstes gliedern wir die Erarbeitung des Projektes in drei Mailsteine, welche die zu erreichenden Ziele im laufe des Projektes beinhalten:

- Meilstein 1
  - Eingewöhnung Knowledge Graphen
  - Eingewöhnung Neo4j und Cyper-Sprache
  - Eingewöhnung Sparql
- Meilstein 2
  - Vorstellung des zu implementierenden Projektes

  - Eingewöhnung Dbpedia und Strukturierung der Daten in der WWW
  - Eingewöhnung OWL, RFD, RFDS
  - Anfragen an die strukturierten Internetwebseiten über die Sparql-Interface
- Meilstein 3
  - Speicherung der erhaltenen Daten
  - Automatizierung der Sparqlanfragen über Java.
  - Automazierte Erstellung des Datenbakes über Java.
  - Erstellung einer Schnittstelle(GUI) für die Benutzer um Anfragen an die Datenbank zu ermöglichen

4 Umsetzung

Die bearbeitung der einzelnen Ziele werden demnächst näher erläutert.

4.1 Knolwledge Graphen

knowledge Graph wurde im 1982 von Hoede and Stokman entwickelt. Sie hatten den Ziel Wissen aus medizinischen und soziologischen Texte zu extrahieren um Expertensysteme zu erhalten und dadurch neue Methoden um das Wissen zu repräsentieren. Knowledge Graphen gehört zu der Kategorie semantischen Netze. [4]

Wissen kommt im Netzwek im strukturierter als auch im unstrukturierter Form vor. Außerdem kann das in deutlicher als auch undeutlicher Form gestaltet sein. Das fürht dazu, dass der nach Information suchenden Netzwerknutzer zu sowohl sinnvollem als auch unbrauchbarem Wissen kommt. Um der Neztwerknutzer einen sinnvollen Überblick über das Wissen im Nezt zu erschaffen, Wird das in sctrukturierter Form aufgebaut. Auf diese Weise kommen wir zur Knowledge Graphen. Was nicht anderes ist als eine Art von semantischen Netzwerke, wo aus der Analysephase von texten zu einer Liste von Konzepte führt. Aus diese Listen werden bennaten Knoten erzeugt und letztendlich verbindet man sie anhand von Relationen, welche die Beziehungen zwieschen Konzepte(Knoten) darstellt und zu der Strukturierung des Wissen hinführt.[2]

4.2 Neo4j und Cyper

Neo4j ist ein open-source NoSQL graph Database, welche im Java und Scala implementiert wurde und von Neo Technology unterstützt wird. Die entwicklug began im 2003 und der freie Zugang ist seit 2007 veröffentlicht worden. Diese Technologien findet Anwendungen im folgenden Bereiche: Matchmaking, Netzwerk-Management, Software-Analytik, wissenschaftliche Forschung , soziale Netzwerke und mehr. Um die Arbeitsweise von Neo4j zu verstehen brauchen wir die Begriffe "Property Graph Databaseünd "NoSQL"kennen lernnen. Ein Property Graph ist, wie der Name einwenig verratet, ein

aus Knoten und kanten Bestehendes Graph.welsches sich von einem normalen Graph dadurch unterscheidet, dass sowohl Knoten als auch Kanten Eigenschaften besitzen können. Und der Begriff "NoSQlïst die Abkürzung für No Only SQL,diese Spracheanfrage umgeht die Leistrungproblemen bei der Anfrage an relationalen Datenbanken in dem durch vernachlässigen von Consistency eine Verbeserung von Availability und Partitioning statt findet. Das fürht dazu, dass die NoSQL Database sich besser skalieren lassen.(Die Abbildung 1 zeigt wie Daten mit Neo4j dargestellt werden

Cypher ist eine deklarative Sprache, welche von Neo4j benutzt wird, Um Anfragen an die Datenbank zu ermöglichen. Ein wichtiger Vorteil ist, dass beim Stellen einer Frage an die Datenbank nur angegeben wird, wonach gesucht wird und nicht wie man es finden soll.[1]

Die Abildung 1 zeigt ein klienes Beispiel,wie Daten im Neo4j dargestellt werden und wie man mit der untenligenden Cypher-Anfrage auf die Knoten und die dazwischen liegende Relation zugreifen kann. Beide Knoten sind von Type City Und die Relation ist von Type Distance

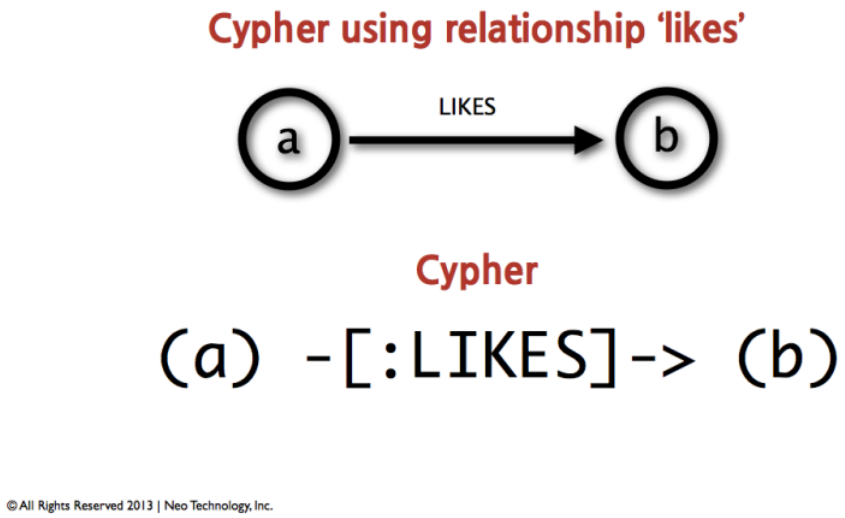

Abbildung 1: Query Beispeil [1]

4.3 Sparql

Sparql ist eine für relationale Datenbanken Anfragesprache, Welche im erste linie entwickelt worden ist um RDF Graphen anzufragen. Der Baustein für SPARQL Abfragen ist Basic Graph Patterns(BGP). SPARQL(BGP) ist eine Menge aus triple patterns, welche im Abschnitt RDF näher erläutert wird. Um komplexe Anfraguen aufzubauen, benutzt man die Operatoren Select, Optional, Union, und Filter. [3]

4.4 Projekt Vorstellung

4.5 Dbpedia und Strukturierung der Daten in der World Wide Web

Dbpedia bietet die von Wikipedia extrahierrte Informationen auf einer strukturierte Weise. Je nachdem was die Anfrage erfordert werden dann die passenden Daten ausgegeben.

4.6 OWL, RDF, RDFS

OWL, RDF und RFDS sind alle Ontologie Sprachen, die dazu dienen eine Ontologie zu beschreiben. Zuallererst erläutern wir kurz was unter dem Begriff Ontologie zu verstehen ist. Dieser bedeutet eine Wissenvermittlung, die eine Bestimmte Dömäne beschreibt. [**?**] Ontology ist bei uns ein Bestandteil von DBpedia da diese die Komplexität von den Anfragen entgegenkommt und lässt die trotz der Komlexität genau beantworten. DBpedia Ontology umfasst Klassen, Eigenschaften und Instanzen.

4.7 Anfragen an die strukturierten Internetwebseiten über die Sparql-Interface

Um Daten aus Internetseiten wie DBpedia zu extrahieren benutzen wir die Snorql Interface, welche unter der LInk http://dbpedia.org/snorql zu erreichen ist. Abbildung 2 zeigt eine Sparql-Anfrague, welche die einige staädte von Deuschland angibt.

4.8 Speicherung der erhaltenen Daten

4.9 Automatizierung der Sparqlanfragen über Java

4.10 Benutzerschnittstelle(GUI)

5 Anleitung

Dieser Abschnitt gibt einen kurzen Überblick über die Vorgehensweise beim Nutzen der entwickelten Knowledgegraphs und des implementierten Codes. Genauere Erläuterungen zu deneinzelnen Abschnitten befinden sich im Kapitel 4.

6 Evaluation

Die im Abschnitt 3 gesetzten Ziele wurden erfüllt. Mit Hilfe der DBpedia und den Sparql Anfragen Deutsche Städte Namen und verschiedene dazugehörige Tourism Informationen. Mitteles eines Java Codes und Neo4j wurden diese erhaltene Informationen, die Tourism in Deutschland inkorporieren im Form von Graphen dargestellt. Anschließend kann der Nutzer anhand eines Interfaces oder auch ohne nützliche gebrauchte Informationen Abragen. Er kann sowohl kürzester Strecke Information als auch Attraktionen einer Stadt abfragen. Die Nutzung DBpedia erleichtet die Arbeit enorm, allerdings sind die Informationen nicht unbedingt Vollständig gespeichert wodurch Probleme entstanden

sind vorallem beim Städten auflisten und beim Verknüpfen von dieser mit den entsprechenden Sehenswürdigkeiten. Das entwickelte Java-Programm wurde unter Windows mit Java 7getestet.

7 Zusammenfassung

In der vorliegenden Arbeit wurde mit Hilfe der Neo4j Graphendatenbank durch Tourism in Deutschland zusammenhängende Informationen ein Knowledgegraph generiert. Anfragen an dieser Graph wurden mit Hilfe der Cypher Anfragesprache realisiert. Die Realisation erfolgte durch DBpedia basiertes Sparql Anfragen und einem Java-Modul, welches im Laufe des Prjekts implementiert wurde.

Literatur

[1] Neo4j. cypher-query-language.

[2] Roel Popping. Knowledge graphs and network text analysis, 2003.

[3] Evren Sirin and Bijan Parsia. Sparql-dl: Sparql query for owl-dl. In *OWLED*, volume 258, page 2, 2007.

[4] Lei Zhang. *Knowledge graph theory and structural parsing*. Twente University Press, 2002.