12 wöchiges Praktikum beim Deutschen Forschungszentrum für Künstliche Intelligenz

Forschungsgruppe Sprachtechnologien (DFKI-LT)
Projekt "Smart Data for Mobility" (SD4M)

DFKI Projektbüro Berlin
Alt-Moabit 91c
10559 Berlin

Praxisbericht

Tom Oberhauser

Matrikelnummer 798158
Studiengang Bachelor Medieninformatik
Beuth Hochschule für Technik Berlin

E-Mail: tom@devfoo.de

Betreuer Prof. Christoph Knabe
Fachbereich VI - Informatik und Medien
Beuth Hochschule für Technik Berlin

Dr. Philippe Thomas
Forschungsgruppe Sprachtechnologie (DFKI-LT)
Deutsches Forschungszentrum für Künstliche Intelligenz

August 26, 2015

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Vorstellung des Praktikumsbetriebes 1

1.2 Weg zur Praktikumsstelle 1

2 Tätigkeitsbereiche und Aufgaben **3**

2.1 Überblick 3

2.1.1 Das Projekt SD4M 3

2.2 Vorbereitung 4

2.3 Aufgaben 4

2.3.1 Extraktion einer Straßenliste aus OpenStreetMap 5

2.3.2 Verknüpfung von Daten der Deutschen Bahn mit Daten aus OpenStreetMap 6

2.3.3 Datenaufwertung?? 6

3 Fazit **7**

3.1 Praktikum und Studium 7

3.2 Bewertung des Praktikums 7

A Anlagen **8**

A.1 Well-Known-Binary Format (WKB) 8

A.2 GeoJSON 9

A.3 OpenStreetMap Datenformat 9

A.4 Beispiel einer Straße in OpenStreetMap 9

Literaturverzeichnis **10**

1 Einleitung

1.1 Vorstellung des Praktikumsbetriebes

Das Deutsche Forschungszentrum für Künstliche Intelligenz GmbH, im folgenden DFKI genannt, wurde 1988 gegründet. Es unterhält Standorte in Kaiserslautern, Saarbrücken, Bremen und ein Projektbüro in Berlin. Mit seinen 478 Mitarbeiten sowie 337 studentischen Mitarbeiten erforscht und entwickelt das DFKI innovative Softwaretechnologien auf der Basis von Methoden der Künstlichen Intelligenz. Die notwendigen Gelder werden durch Ausschreibungen öffentlicher Fördermittelgeber wie der Europäischen Union, dem Bundesministerium für Bildung und Forschung (BMBF), dem Bundesministerium für Wirtschaft und Technologie (BMWi), den Bundesländern und der Deutschen Forschungsgemeinschaft (DFG) sowie durch Entwicklungsaufträge aus der Industrie akquiriert.[1]

Ich absolvierte mein Praktikum innerhalb der *Forschungsgruppe Sprachtechnologie*, einer von 15 Forschungsgruppen[2] des DFKI, im Projektbüro Berlin. Die Gruppe wird geleitet durch Prof. Dr. Hans Uszkoreit.[3].

Meine Aufgabengebiete konzentrierten sich um das Projekt *"SD4M - Smart Data for Mobility"*. Das DFKI ist hier Teil eines Konsortiums aus 5 Partnern unter der Konsortialführung der *DB Systel GmbH*.[4]. Das Projekt *SD4M* wird in Abschnitt 2.1.1 auf Seite 3 näher erläutert.

1.2 Weg zur Praktikumsstelle

Herr Prof. Dr. habil. Alexander Löser aus dem Fachbereich VI der Beuth Hochschule für Technik Berlin machte mich auf den Praktikumsplatz aufmerksam. Durch seine Mitarbeit in Projekten im DFKI Projektbüro Berlin hatte er wargenommen, dass Bedarf und Interesse an Praktikanten und studentischen Mitarbeitern besteht und mich benachrichtigt. Ich habe mich daraufhin auf der Website des DFKI über die aktuellen Projekte informiert. Da ich mich sehr für Datenintegration interessiere und

[1] http://www.dfki.de/web/ueber
[2] http://www.dfki.de/web/ueber/orgaeinheiten
[3] http://www.dfki.de/lt/
[4] http://sd4m.net/konsortium

ich vor meinem Studium bereits Berufserfahrung auf diesem Gebiet gesammelt habe, fand ich das Projekt *SD4m - Smart Data for Mobility* sehr interessant. Es umfasst zwei Themenkomplexe: zum einen die Verknüpfung unterschiedlicher Datenquellen und zum anderen Methoden des *Natural Language Processings* bzw. des *Text Minings*. Das Thema Text Mining wurde kurz im Modul Datenbanksysteme im zweiten Semester angeschnitten und es hatte mich auch sehr interessiert. Nach einem persönlichen mit Herr Uszkoreit, in dem ich mein Interesse für das Projekt darlegen konnte, kam es zur Vertragsunterzeichnung.

2 Tätigkeitsbereiche und Aufgaben

2.1 Überblick

Ich habe im Rahmen meines Praktikums am Projekt *SD4M - Smart Data for Mobility* mitgearbeitet. Meine konkrete Aufgabe war die Aufbereitung und Integration verschiedener Daten und Datenbanken, damit diese im Projekt Verwendung finden können. Meine Hauptdatenquelle waren die Geodaten des OpenStreetMap[1] Projekts.

2.1.1 Das Projekt SD4M

Das Projekt *Smart Data for Mobilty*[2], im folgenden *SD4M* genannt, ist ein Verbundprojekt eines Konsortiums aus 5 Partnern und wird vom Bundesministerium für Wirtschaft und Energie gefördert. Das Konsortium besteht aus 4 Wirtschaftsunternehmen und dem DFKI als Forschungseinrichtung.

- DB Systel GmbH (Konsortialführung)
- Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
- idalab GmbH
-]init[AG für digitale Kommunikation
- PS-Team Deutschland GmbH Co. KG

Ziel des SD4M Projekts ist eine branchenübergreifende Serviceplattform, welche Daten der unterschiedlichen Mobilitätsanbieter (z.B. der Fahrplan der Deutschen Bahn) sowie öffentliche verfügbare strukturierte und unstrukturierte Daten (z.B. Twitter oder Facebook) miteinander verknüpft. Diese verknüpften Daten sind für Endnutzer, aber auch für Unternehmen oder die öffentliche Verwaltung von Interesse. In Abbildung 1 wird verdeutlicht, wie sich aus unstrukturierten Twitter-Daten Verspätungsinformationen für konkrete Verkehrsmittel extrahieren lassen. Diese können dann Endnutzern oder den Mobilitätsanbietern zur Verfügung gestellt werden.

[1] http://www.openstreetmap.org/
[2] http://sd4m.net/

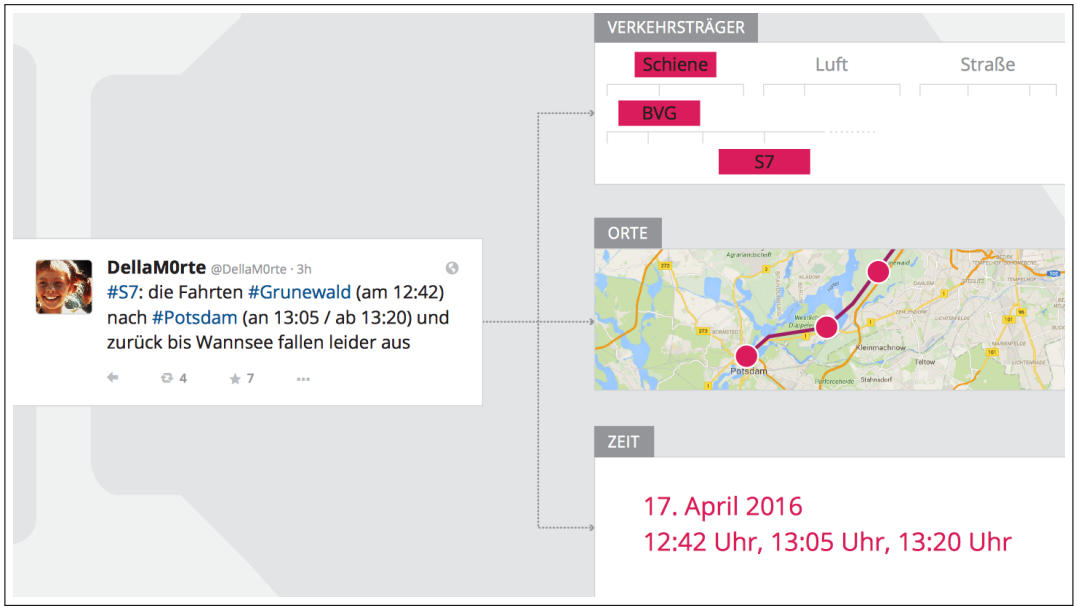

Abb. 1.: Verknüpfung eines Tweets mit Fahrplandaten[@Ing16]

2.2 Vorbereitung

Beim ersten Gespräch mit meinem Praktikumsbetreuer Dr. Philippe Thomas informierte ich mich, welche Programmierumgebungen und Programmiersprachen beim DFKI üblich sind. Ebenfalls erkundigte ich mich nach einer vorhandenen OpenStreetMap Datenbank und weiterer vorhandener Infrastruktur. Wir klärten, dass Java die geeignetste Sprache zur Lösung meiner Aufgaben war. Ebenfalls war eine OpenStreetMap Datenbank mit dem Datenbestand von Deutschland, sowie ein GitLab Repository Server vorhanden. Da ich auf meinem eigenen Notebook entwickeln wollte, installierte ich mir einen virtuellen Linux-Server mit einer PostgreSQL Datenbank um einen kleinen Teil der OpenStreetMap Daten lokal auf meinem Rechner zu haben. So konnte ich schneller entwickeln und mit einer wesentlich kleineren Datenbank testen.

2.3 Aufgaben

Meine Aufgaben während des Praktikums lassen sich in 3 Teilbereiche gliedern. Zunächst sollte ich eine Straßenliste aus OpenStreetMap extrahieren. Anschließend verknüpfte und aggregierte ich Daten, welche von der Deutschen Bahn geliefert wurden, mit Daten aus OpenStreetMap. In den letzten Wochen meiner Praxisphase gab es dann noch verschiedene Datenaufbereitungsaufgaben. Auf diese 3 Themengebiete werden in diesem Abschnitt nun eingegangen.

2.3.1 Extraktion einer Straßenliste aus OpenStreetMap

Aufgabe

Meine erste Aufgabe bestand darin, eine Liste aller Straßen Deutschlands inclusive dazugehöriger Geodaten zu erstellen. Es sollte eine Java Anwendung erstellt werden, welche via Kommandozeilenargumenten konfiguriert wird, und die entsprechenden Ergebnisse in einer CSV-Datei ablegt. Im Ergebnis sollten pro zusammenhängender Straße die Daten

- **ID**, eine fortlaufende Nummer
- **Name**, der Name der Straße
- **LineString**, der geographische Straßenverlauf im *Well-Known-Binary (WKB)* Format (siehe Anlage A.1)
- **GeoJSON**, der geographische Straßenverlauf im *GeoJSON* Format (siehe Anlage A.2)

vorhanden sein. Diese Daten sollen anschließend zur geographischen Verortung von Straßen aus Tweets genutzt werden.

Lösung

Für meinen Anwendungsfall, die Filterung und Extraktion von Daten, war es notwendig, die OpenStreetMap Daten in einer Datenbank vorliegen zu haben. Der Hauptgrund hierfür ist dem Datenformat der OpenStreetMap Daten geschuldet. Wie im Anhang A.3) aufgezeigt, beinhalten die OpenStreetMap Daten *Nodes* (Punkte mit Koordinaten) und *Ways* (Linien aus Punkten).

Eine physikalisch vorhandene Straße ist wird durch mehrere aneinanderhängende *Ways* repräsentiert. Wenn sich im Straßenverlauf ein Attribut (Ein *Tag*) der Straße ändert, zum Beispiel ein hinzukommender Radweg, muss ein neues Teilstück diese neuen Gegebenheiten abbilden. Ein Beispiel hierfür findet sich in Anlage A.4.

> "osm2pgsql is mainly written for rendering data with data. So it only imports tags which are going to be useful for rendering. ... whereas osmosis and osmium more geared towards truthfully representing a full OSM data set."[3]

[3] http://giswiki.hsr.ch/Osm2pgsql

Schwierigkeiten

Das Testen auf einer kleinen Datenbank beschleunigte zwar die Entwicklung, aber erst der erste Test auf der Hauptdatenbank zeigte massive Performanceprobleme auf. Ich hatte wollte meine Anwendung möglichst speicherschonend gestalten und holte
... meine Straßenobjekte einzeln aus der Datenbank...
Eine Messung zeigte jedoch dass die Datenbankabfrage der Flaschenhals
... mit 300ms Abfragezeit und unter 1ms Verarbeitungszeit
... also neugeschrieben und alles initial in den RAM zum verarbeiten...

2.3.2 Verknüpfung von Daten der Deutschen Bahn mit Daten aus OpenStreetMap

Aufgabe

Lösung

Schwierigkeiten

2.3.3 Datenaufwertung??

3 Fazit

3.1 Praktikum und Studium

3.2 Bewertung des Praktikums

Anlagen

A

A.1 Well-Known-Binary Format (WKB)

Das *Well-Known-Binary* Format ist die binäre Repräsentation eines geometrischen Objekts des *Simple Feature Models*. Das Simple Feature Model ist eine Untermenge des ISO 19107 Standards, welcher die geometrischen Eigenschaften von Geoobjekten spezifiziert. Ausgehend von einer allgemeinen Oberklasse können geometrische Primitive, wie z.B. ein Punkt, oder komplexe geometrische Objekte, wie z.B. Flächen oder Sammlungen von Objekten, beschrieben werden. (vgl. [Bil10]:358ff.). Die verfügbaren Klassen werden in Abbildung 2 aufgezeigt.

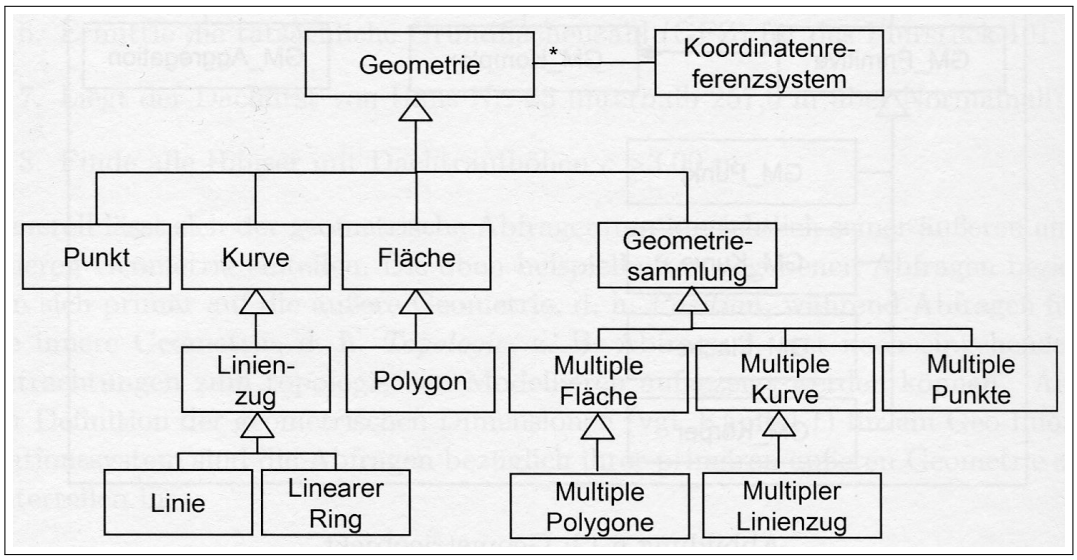

Abb. 2.: Geometrien im Simple Feature Model [Bil10]:360

Das WKB Format wird beispielsweise innerhalb der PostgreSQL Erweiterung PostGIS[1] genutzt, um geometrische Objekte in einer Datenbank abzulegen. Analog dazu existiert das *Well-Known-Text* Format, welches die textuelle Repräsentation geometrische Objekte des Simple Feature Models spezifiziert.

[1] http://postgis.net

Beispiel für das WKB und das WKT Format

Ein Punkt mit den Koordinaten `13.439561 Ost` sowie `52.54002 Nord` entspricht der WKT Repräsentation

```
SRID=4326;POINT(13.439561 52.54002)
```

und der WKB Repräsentation

```
0101000020E6100000B131AF230EE12A40CC9717601F454A40
```

Der Wert `SRID` beinhaltet die ID des zu verwendenden Koordinatenreferenzsystems. Die ID `4326` entspricht dem häufig verwendeten Referenzsystem `WGS84`[2].

A.2 GeoJSON

GeoJSON[3] ist ein Format zur Repräsentation von geometrischen Objekten im JSON Format. Es verwendet ein ähnliches hierarchisches Klassenmodell. (vgl. [@How08]).

Beispiel für das GeoJSON Format

Ein Punkt mit den Koordinaten `13.439561 Ost` sowie `52.54002 Nord` entspricht der GeoJSON Repräsentation

```
{
  "type": "Point",
  "coordinates": [
      13.439561,
      52.54002
  ]
}
```

A.3 OpenStreetMap Datenformat

A.4 Beispiel einer Straße in OpenStreetMap

[2] http://spatialreference.org/ref/epsg/4326/
[3] http://geojson.org

Literaturverzeichnis

[Bil10] Ralf Bill. *Grundlagen der Geo-Informationssysteme*. 5., völlig neu bearb. Aufl. Berlin [u.a.]: Wichmann, 2010 (zitiert auf Seite 8).

Online-Quellen

[@How08] Howard Butler (Hobu Inc.), Martin Daly (Cadcorp), Allan Doyle (MIT), Sean Gillies (UNC-Chapel Hill), Tim Schaub (OpenGeo), Christopher Schmidt (MetaCarta). *The GeoJSON Format Specification*. 2008. URL: `http://geojson.org/geojson-spec.html` (besucht am 23. Mai 2016) (zitiert auf Seite 9).

[@Ing16] Ingo Schwarzer. *Smart Data For Mobility (SD4M) – Projekt-Präsentation*. 2016. URL: `http://www.sd4m.net/sites/default/files/publications/%20SD4M-Pr%C3%A4sentation.pdf` (besucht am 10. Mai 2016) (zitiert auf Seite 4).

Declaration

You can put your declaration here, to declare that you have completed your work solely and only with the help of the references you mentioned.

Berlin, August 26, 2015

Tom Oberhauser