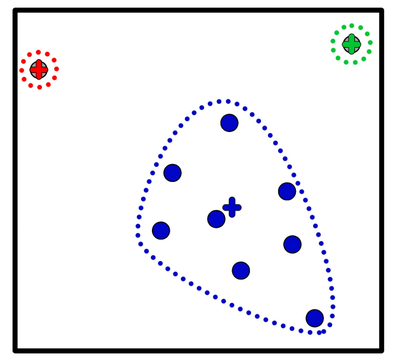
Fachhochschule Brandenburg
Fachbereich Informatik und Medien

Entwicklung und Evaluierung von Clustering-Verfahren für Points of Interest verschiedener thematischer Kategorien



Bachelorarbeit, vorgelegt von Bertram Sändig

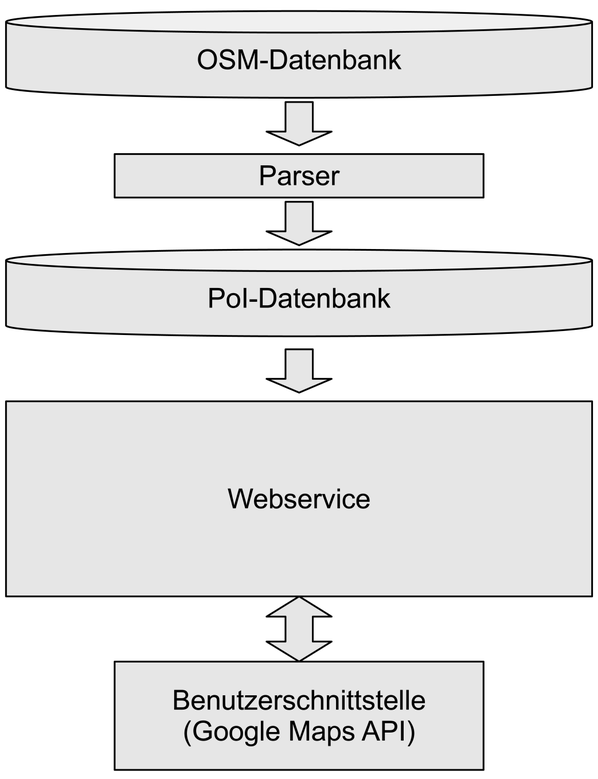
Aufgabenstellung

In dieser Arbeit geht es um die Entwicklung einer Anwendung zur Darstellung von konzentrierten Gebieten von Points of Interest (PoIs). Dadurch soll einem Anwender z. B. ein Ballungsgebiet von Restaurants innerhalb des urbanen Umfelds angezeigt werden.

Der Vorgang, Wertemengen nach ihrer Ähnlichkeit (hier: die geographische Distanz) zu gruppieren, ist eine Disziplin des Data-Minings und wird als Cluster-Analyse bezeichnet.

Verschiedene Clustering-Verfahren sollen getestet und angepasst werden, um ihre Eignung für die Aufgabenstellung zu bewerten und zu optimieren.

Systemarchitektur



Ein in Java geschriebener SAX-Parser entnimmt alle für die Anwendung relevanten PoI-Daten aus der freien OpenStreetMap-Datenbank und legt sie in einer SQL-Datenbank ab, die für die Anwendung erstellt wurde.

Ein RESTful Web Service (dessen Methoden ebenfalls in Java geschrieben wurden) entnimmt angefragte Daten aus der SQL-Datenbank, führt darauf Clustering-Operationen aus und gibt das Ergebnis im JSON-Format zurück.

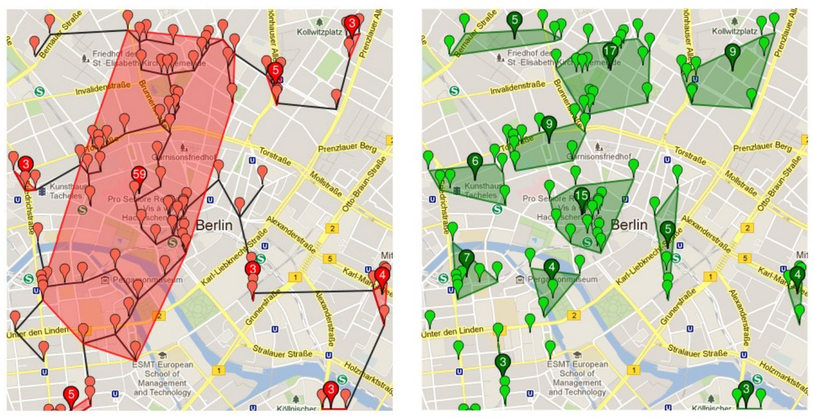
Die Benutzerschnittstelle wurde mithilfe der Google Maps JavaScript API Version 3 umgesetzt. Sie macht für den Benutzer die gewünschten Anfragen an den Webservice und visualisiert die Ergebnisse auf der Karte.

K-Means Clustering

Bei diesem Verfahren wird versucht, Cluster zu bilden, bei denen alle Punkte möglichst nahe am Zentrum (Schwerpunkt) des Clusters liegen. Dies wird erreicht, indem zuerst zufällige Zentren gewählt werden. Dann werden abwechselnd die Punkte den nächstliegenden Zentren zugeordnet und daraufhin die Zentren neu berechnet, bis sich nichts mehr ändert.

Minimum spanning tree Clustering

Bei diesem Verfahren wird zuerst der minimale Spannbaum (kürzester Teilgraph der alle Knoten des Gesamtgraphen enthält) gebildet. Die Standardmethode, um mithilfe dieses minimum spanning (MST) tree eine Gruppe von *k* Clustern zu erzeugen, ist es, die Kanten des Baums absteigend nach ihrer Länge zu sortieren, um daraufhin die *k* - 1 längsten Kanten zu entfernen.



Vergleich MST-Clustering (l.) und K-Means-Clustering (r.)

DK-Means-Algorithmus

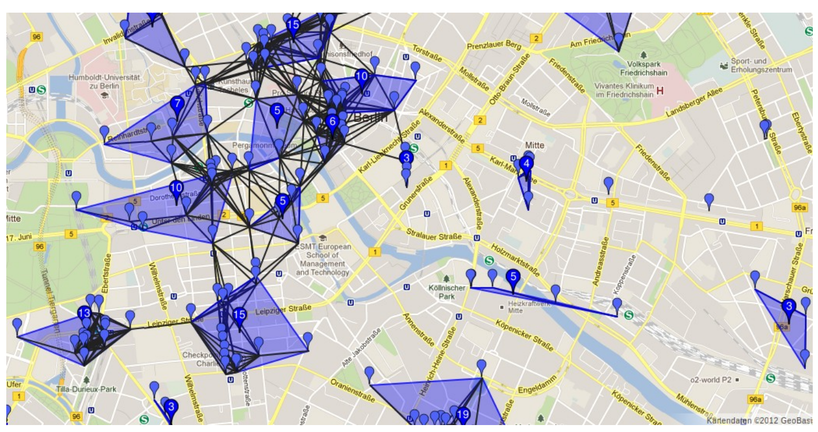
Der dk-Means (gesprochen wie „decay“, engl. „Zerfall“, steht für dual k-Means oder distance k-Means) oder auch dtk-means (distance threshold k-Means) ist eine Erweiterung des k-Means Algorithmus die speziell für diese Arbeit entwickelt wurde um dem Bedürfnis der Aufgabenstellung zu entsprechen, dass Cluster kompakt sein sollen und Punkte innerhalb eines Clusters eine bestimmte Distanz zueinander nicht überschreiten dürfen.

Phase 1:

In der ersten Phase wird aus der Punktemenge ein Graph gebildet indem jeder Punkt mit allen anderen Punkten verbunden wird, deren Entfernungen zu ihm unter einem bestimmten Schwellenwert (maximale Lauftoleranz) liegen.

Phase 2:

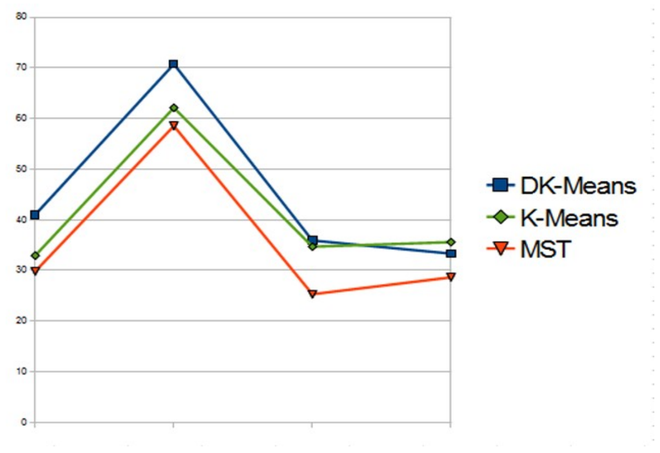
In Phase zwei wird die Fläche der entstandenen Cluster berechnet und mit einem zweiten Schwellenwert verglichen (gewünschte Cluster-Größe). Auf alle Cluster die größer sind, wird der k-Means Algorithmus angewandt, um zu große (möglicherweise Kettenartige) Cluster zu verhindern.



Testreihe 1 Mst-, K-Means-, DK-Means-Clustering (v. l. n. r.)

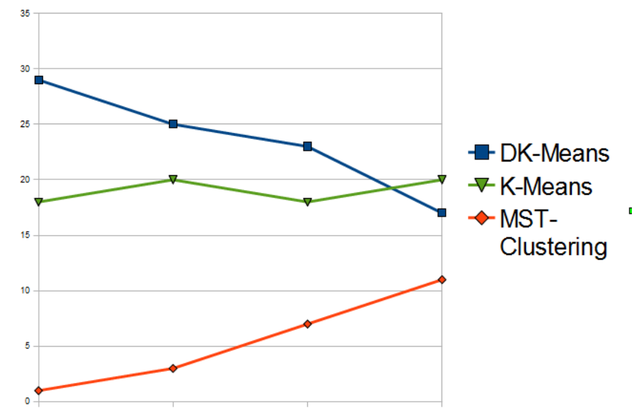
Auswertung

Die Verfahren wurden verglichen, indem vier verschiedene Eingabemengen mit den drei Methoden geclustert wurden. Daraufhin wurde die Dichte der entstandenen Ergebnismengen berechnet, die definiert ist als die Menge an PoIs je km².



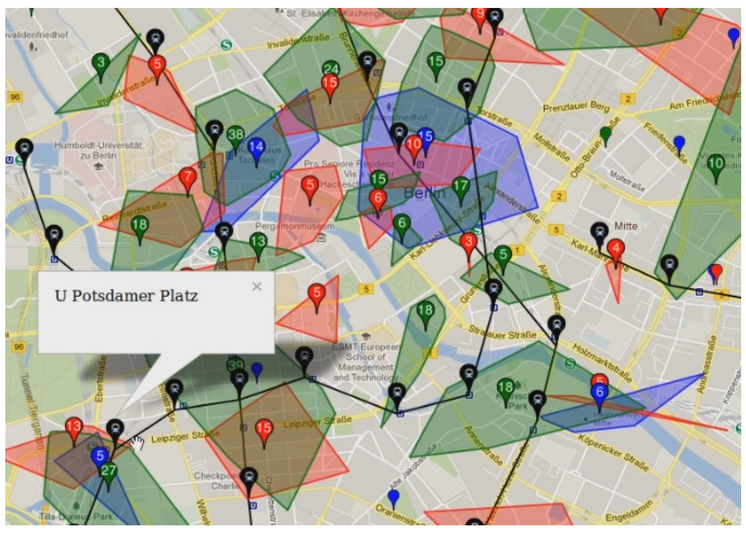
Vergleich Dichte in PoIs je km²

Da die Beewertung von Clustering-Ergebnissen subjektive Eigenschaften besitzt, wurde im zweiten Teil der Evaluierung eine Gruppe von 15 Testpersonen befragt welches Ergebnis ihnen am meisten zusagt.



Umfrageergebnisse

Ergebnis



3 Cluster-Kategorien (blau = bar, rot = cafe, grün = restaurant) und U-Bahnlinien (schwarzes Netz)

Betreuer: Dipl.-Inform. I. Boersch
Dr. I. Seifert

Abgabe: 25.01.2012