Anwendung von Frequent Itemset Mining auf nutzergenerierte Geodaten

Christian Sengstock
Universität Heidelberg
Institut für Informatik
Lehrstuhl für Datenbanksysteme
sengstock@informatik.uni-heidelberg.de

Michael Gertz
Universität Heidelberg
Institut für Informatik
Lehrstuhl für Datenbanksysteme
gertz@informatik.uni-heidelberg.de

1. EINFÜHRUNG

Geodaten werden zunehmend durch Nutzer generiert. Openstreetmap (OSM), Google Earth, Qype sowie mit geographischen Attributen annotierte Nutzerinhalte wie Wikipedia-Artikel, Blogs und Tweets sind einige Beispiele. Die Datenformate und -modelle sind oft standardisiert (GML, KML) oder es handelt sich um einfache flache Modelle (GeoJSON, OSM-Format, zeilenbasierte Formate). Die Modelle erlauben dabei die freie Definition von Attributen (Metadaten, Sachdaten) in unterschiedlicher Mächtigkeit.

Wir stellen einen Ansatz und ein Framework vor, das die statistische Analyse, das Data Mining und die Integration solcher Datenmengen vereinfacht. Hierzu werden die Daten in ein flaches *Key-Value-Modell* überführt, wobei jedes *Key-Value-Paar* eine *Eigenschaft* darstellt. Auf Basis dieser *eigenschaftsbasierten* Sichtweise lassen sich Verfahren des *Frequent Itemset Mining* auf die Daten anwenden, um automatisiert Schema-Modelle zu identifizieren und zu analysieren.

Dieser Ansatz bildet eine Grundlage für Anwendungen und Verfahren wie Qualitätsanalyse, Schema-Integration, Clustering und Klassifikation, Visualisierung und die automatisierte Generierung von Ontologien.

In dieser Arbeit verwenden wir den genannten Ansatz, um die Qualität und die Konsistenz der frei editierbaren OSM-Daten automatisiert zu analylsieren. Hierzu wird auf die ermittelten Schemata eine Konsistenzprüfung der Wertebereiche durchgeführt sowie ein Qualitätsmaß berechnet und auf Karten visualisiert.

2. VERWANDTE FORSCHUNGSTHEMEN

Die Analyse der Eigenschaften von Geodaten wird auf unterschiedlichen Ebenen untersucht. Im *frequent geographic pattern mining* wird das Auftreten von geographischen Mustern auf Basis geographischer Eigenschaften (*contains, intersects, near_by*) analysiert [2]. Suchmöglichkeiten auf Basis semantisch-geographischer Eigenschaften sind eine Forschungsfrage in Bezug auf ein *Geospatial Semantic Web* [3]. Speziell mit der Analyse und Bewertung von nutzergenerierten Annotationen im Rahmen von Verschlagwortungen haben sich Golder et al. [4] beschäftigt. In unserer Arbeit konzentrieren wir uns dagegen nicht auf die semantische Analyse, sondern auf Formen der automatisierten Schemaanalyse, um gemeinsam verwendete Schemata und Konzeptualisierungen in von Nutzern generierten Geodaten zu finden.

3. FREQUENT ITEMSET MINING

Das *Frequent Itemset Mining* wurde Anfang der 90er Jahre entwickelt, um Zusammenhänge von großen Warenkorb Datensätzen zu analysieren [1]. Hierbei wird von einer Transaktion ausgegangen (typischerweise eine Kauftransaktion), der eine Menge von Waren (*Items*) zugeordnet sind. Durch das Key-Value-Modell kann man ein Objekt auf das Transaktions-Waren-Modell abbilden. Das Objekt (identifiziert durch eine ObjektID) ist die Transaktion. Die Key-Value-Paare entsprechen den zugeordneten Waren.

Um das im Folgenden beschriebene Frequent Itemset Mining auf Geodaten anwenden zu können, wird jedes Geoobjekt, bestehend aus einer Geometrie und einer Menge an Metadaten in ein Transaktionsobjekt mit zugeordneter Geometrie und einer Menge aus Key-Value-Paaren (Eigenschaften) überführt:

```
Geo-Transaktionsobjekt(Geometrie, Liste((Key,Value)))
```

Die Abbildung des *Frequent Itemset Mining* auf die Attribute von Datenmodellen nennen wir *Frequent Property Mining*, eine gefundene häufige Menge von Eigenschaften ein *Frequent Property Set*.

Um häufig auftretende Mengen von Key-Value-Paaren zu ermitteln, wird ein *Frequent Itemset Mining* auf die Geo-Transaktionsobjekte mit den zugeordneten Key-Werten ausgeführt. Hierfür haben wir einen leicht angepassten *Apriori-Algorithmus* [1] verwendet. Das Ergebnis sind Mengen aus Keys (Frequent Property Set), wobei jeder Menge seine Häufigkeit (Support) im Datensatz D zugeordnet ist ($Anzahl/|D|$). Beispielsweise:

```
(highway) 0.3
(highway name) 0.25
(highway access) 0.2
(highway access name) 0.18
```

Die zu einem Frequent Property Set gehörenden Geoobjekte lassen sich als individuelle Datenmengen weiterverarbeiten, visualisieren und exportieren. Für das im Folgenden vorgestel- lte Framework bildet die Generierung der Frequent Property Sets die Grundlage für weitere Analysen.

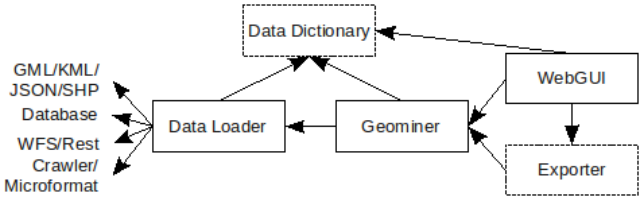

Abb. 1: Komponenten des Frameworks

4. FRAMEWORK KOMPONENTEN

Der *Data Loader* (siehe Abb. 1) liest die Geodaten und überführt sie in das Key-Value-Modell. 1:N Beziehungen müssen dabei auf Listen abgebildet werden. Um unterschiedliche Bezeichnungen von Keys zu unterstüzen, können alternative Bezeichner in einem *Data Dictionary* abgelegt werden. Das Data Dictionary kann als Ontologie weiterentwickelt werden, um semantische Beziehungen zwischen den Attributen abbilden zu können.

Der *Geominer* liest die im Key-Value-Modell verfügbaren Daten und führt zunächst ein einfaches Frequent Property Mining durch. Dieses bildet die Basis für weitere Verfahren, wie statistische Analysen (insbesondere Histogramme und Wertebereiche) oder Clustering-Verfahren. Der Geominer kann auf das Data Dictionary zugreifen, um alternative Bezeich-

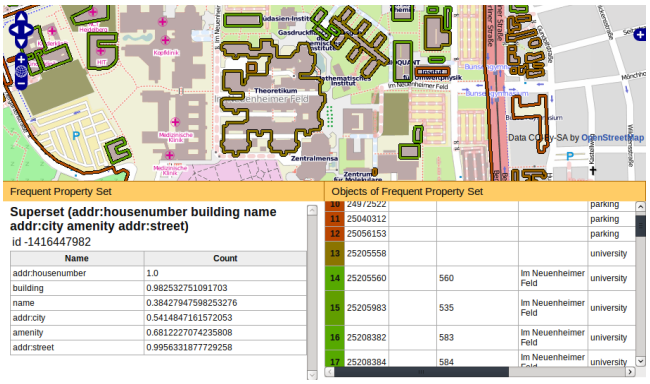

Abb. 2: Bewertetes Frequent Property Set

ner zu erfahren.
Um die Geominer Analysen auszuführen und zu visualisieren, wurde eine *WebGUI* entwickelt. Diese erlaubt es, die zu analysierenden Datenquellen sowie die räumlichen Anfragebereiche auszuwählen und die generierten Frequent Property Sets auf Karten sowie tabellarisch darzustellen. Zudem können die generierten Datensätze in Formaten wie GML oder als Shapefile exportiert werden, um zur weiteren Analyse externe GIS-Software verwenden zu können.
Die Kartendarstellung in der WebGUI erfolgt auf Basis eines dynamischen WMS, welcher von Desktop- und Web-GIS-Systemen als Layer verwendet werden kann.

5. QUALITÄT VON OPENSTREETMAP ATTRIBUTEN

Für dieses Beispiel wurden die OSM-Daten als Datenbasis verwendet. Diese Daten basieren auf einer offenen Zuordnung von Attributen zu den Geoobjekten und sind somit für die Analyse auf Basis von Key-Value-Paaren sehr gut geeignet.
Als Raumausschnitt für die folgenden Beispiele diente Heidelberg, als Geoobjekte wurden die Liniengeometrien mit den zugeordneten Attributen verwendet. Das Frequent Property Mining auf den Liniendaten ergab unter anderem folgende Frequent Property Sets:

```
(building addr:street addr:housenumber) 0.073
(building amenity) 0.047
```

Auf Basis der gefundenen Sets lassen sich Assoziationsregeln erstellen (wenn `building addr:street addr:housenumber => amenity`) und weitere Zusammenfassungen durchführen (siehe unten). Zudem lassen sich durch eine einfache statistische Auswertung der Wertebereiche der einzelnen Sets interessante Aussagen über die Qualität der Eigenschaften treffen. Beispielsweise sind die Wertebereiche für das Set `(building,amenity)`:

```
building -> (yes, university, hospital)
amenity -> (university, hospital)
```

Das gefundene häufige Konzept `(building,amenity)` besitzt in den Wertebereichen der Eigenschaften Redundanzen. Für ein *gutes* Konzept sollten die Wertebereiche jedoch disjunkt voneinander sein. Die Frequent Property Sets können nach dem Grad der disjunkten Wertebereiche sortiert werden.

Auf Basis einer Zusammenfassung von Frequent Property Sets lassen sich Mengen aus Key-Value-Paaren finden, welche eine bestimmte Domäne geeignet abbilden. Wir haben den in Alg. 1 beschriebenen Algorithmus verwendet, um Frequent Property Sets zu vereinen, die gemeinsame Eigenschaften besitzen. Entsprechend dem Support der Frequent Property Sets wird den Eigenschaften ein Gewicht-Attribut im vereinten Set S zugeordnet. Auf Basis der Eigenschaften $S_{1..n}$ für $n = |S|$ und deren Gewichte $S_i.weight$, lässt sich ein Qualitätsmaß eines Records R mit den Eingenschaften $R_{1..n}$ berechnen (für nicht existierenden Werte R_i ist der Wert 0):

$$\frac{\sum_{i=1}^{n} S_i.weight * R_i}{n}$$

Ein hoher Wert wird erreicht, wenn viele Eigenschaften des Records einen Wert für einen Key des vereinten Sets besitzen. Ein niedriger Wert bedeutet eine schlechte Qualität des Datensatzes, gemäß der Eigenschaften und der Gewichte des vereinten Sets.
Ein Beispiel für ein vereintes Set mit zugeordneten Gewichten ist:

```
( addr:city 1.000, addr:street 0.932, building 0.932,
addr:housenumber 0.909, amenity 0.636, name 0.5 )
```

Die Visualisierung der Records auf Basis des Qualitätsmaßes ist in Abb. 2 dargestellt. Eine weiteres Set, welches einen Großteil von Straßen und Wegen repräsentiert, ist:

```
(highway 1.000, bicycle 0.214, name 0.204, access 0.164, created_by
0.118, oneway 0.102, surface 0.073, service 0.071, maxspeed 0.071,
area 0.037, foot 0.034, layer 0.031, lanes 0.031, tunnel 0.029,
source 0.029, ref 0.026, tracktype 0.025, bridge 0.018,
cycleway 0.013, sac_scale 0.013, maxheight 0.008 )
```

Die gefundenen Sets können für Qualitätsaussagen von Geodaten verwendet und automatisiert aktualisiert werden. Ein praktisches Beispiel wäre etwa das Vorschlagen von geeigneten Eigenschaften und Wertebereichen während des Editierens eines Geoobjekts.

Alg. 1 Berechnung von vereinten Sets

```
fsets = list[all frequent property sets]
usets = dictionary[empty]
for f_a in fsets do
  for f_b in fsets do
    if f_b contains f_a and f_b is not f_a then
      uset = usets[f_a]
      if not exists uset then
        usets[f_a] = f_a
      end if
      for item in f_b \ f_a do
        weight = f_a.support / f_b.support
        uset[f_a].add( (item,weight) )
      end for
    end if
  end for
end for
return usets
```

6. AUSBLICK

Die Verarbeitung von Geodaten in einem vereinfachten Key-Value-Modell sowie die Anwendung von Frequent Itemset Methoden kann als Grundlage für statistische Analysen und geographisches Data Mining dienen.
Unser Interesse liegt in weiteren Analysen und Tests durch die Integration von unterschiedlichen Geodaten. Ein wesentlicher Bereich ist auch die parallele Verarbeitung von geographischen Data Mining Verfahren. Das vorgestellte Framework und die Analysen sollen als Grundlage für ein paralleles Geomining-Framework dienen.

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB '94: Proc of the 20th Int Conf on Very Large Data Bases*, pages 487–499, 1994.

[2] V. Bogorny, S. Camargo, P. M. Engel, and L. O. Alvares. Mining frequent geographic patterns with knowledge constraints. In *GIS '06: Proc of the 14th annual ACM Int Symp on Advances in geographic information systems*, pages 139–146, 2006.

[3] M. J. Egenhofer. Toward the semantic geospatial web. In *GIS '02: Proc of the 10th ACM Int Symp on Advances in geographic information systems*, pages 1–4, 2002.

[4] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.