Exploring Volunteered Geographic Information Using Scale-Dependent Frequent Pattern Mining

Christian Sengstock, Michael Gertz

Institute of Computer Science, University of Heidelberg
Email: {sengstock,gertz}@informatik.uni-heidelberg.de

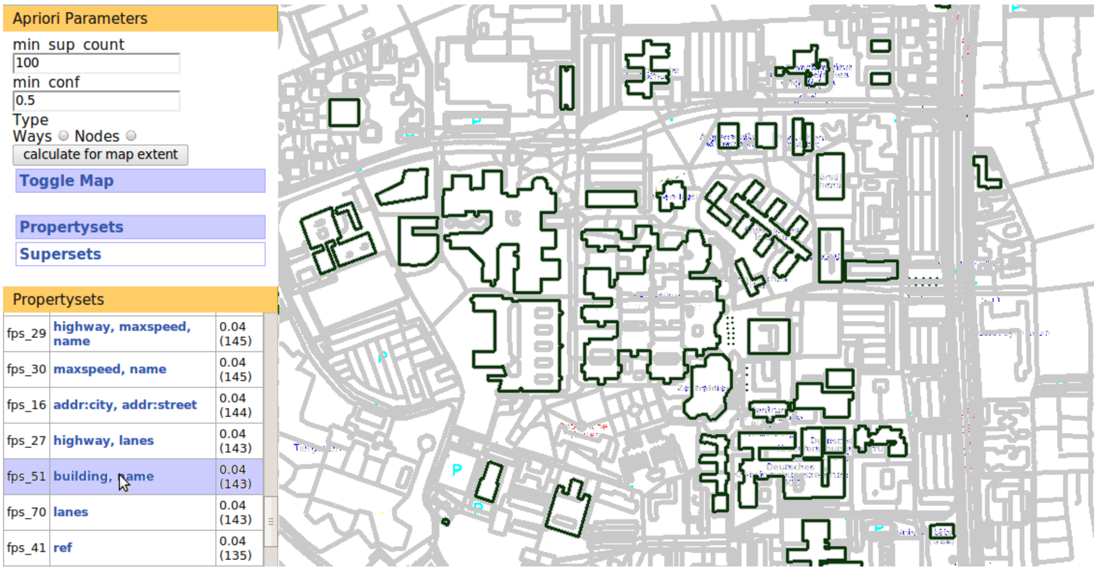
1. Volunteered Geographic Information

There is an explosion of geographic information generated by individuals on the Web. Users provide geotagged photos and tweets, geotag Wikipedia articles, create gazetteer entries, update geographic databases like OpenStreetMap (OSM) and much more. Such user-generated geodata, also called Volunteered Geographic Information, VGI (Goodchild 2007), is becoming an important source for geo-services like map generation, routing, search, spatial analysis and mashups. Different from traditional geodata, VGI often has no distinct classifying attributes or explicit taxonomy. Users are free to create new tagging schemas or add new properties or text. Although some schema checks may exist on the editor level through auto-completion or templates, these checks are not strict and can be ignored by the user.

Analyzing the dynamic and heterogeneous schemas of VGI to find common conceptualizations is an important and complex task. For example, Deng et al. (2009) use density based clustering and a document term matrix to find conceptualizations in geotagged Flickr images. Edwardes and Purves (2007) explore the potential to develop a hierarchy of place concepts based on co-occurring characteristic terms in the description of geotagged photos of the British Isles. Extracting and exploring concepts is an important prerequisite to analyze the quality and consistency of a dataset and to evaluate its “fitness for use” (Gervais 2009).

We describe our work on using frequent pattern mining to extract and explore conceptualizations of VGI. Frequent pattern mining is used for effective classification in association rule mining (Liu et al. 1998). Afrati et al. (2004) use frequent sets to find approximate patterns, which is a promising technique for concept extraction and exploration. For geospatial data, frequent pattern mining is used to determine spatial association rules (Koperski 1995) and to perform co-occurrence analysis (Han 2009). In our approach we transform VGI into a flat model of transaction objects, which can be input to frequent pattern mining algorithms.

Different from transactions of market basket data, which are the typical input to frequent pattern mining, geospatial patterns may occur rarely in a dataset but are nevertheless interesting. Ding et al. (2006) introduce a framework to mine regional association rules based on prior clustering to find patterns in subregions. However, to extract concepts, mining subregions is not an option. We explain what extensions to frequent pattern mining are needed to deal with the scale-dependency and introduce a bottom-up mining approach based on quadtrees. We developed a prototype framework to mine the frequent patterns apriori, which then can be efficiently accessed by clients. For this, we describe the OSM Explorer, which visualizes frequent patterns in the OSM dataset and performs data consistency and quality checks.



min sup count: 100

min conf: 0.5

Type: Ways ○ Nodes ○

Propertysets

fps_29	highway, maxspeed, name	0.04 (145)
fps_30	maxspeed, name	0.04 (145)
fps_16	addr:city, addr:street	0.04 (144)
fps_27	highway, lanes	0.04 (143)
fps_51	building, name	0.04 (143)
fps_70	lanes	0.04 (143)
fps_41	ref	0.04 (135)

Figure 1. OSM Explorer: Visualization of the pattern (building, amenity).

2. Transaction Model

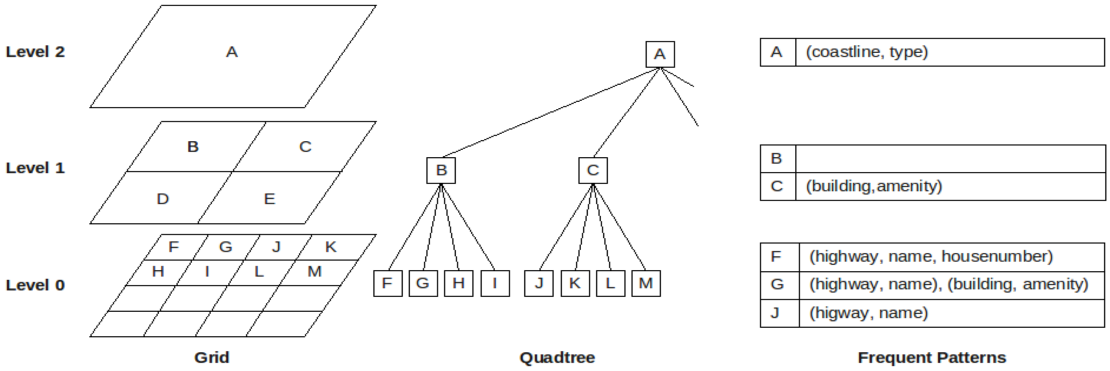
To employ frequent pattern mining to extract concepts from VGI, the heterogeneous geographic information needs to be transformed into transactions. A transaction has an associated set of items and is input record frequent pattern mining. We view each geoobject as a transaction having a geometry and a set of attributes. Attributes can be key-value pairs (representing an attribute name and value) or just keys (like tags). Text has to be itemized first. For example, by using frequency term vectors or by extracting named entities, a text describing geographic information can be transformed into a set of attributes. In general, a geoobject is represented as a transaction as follows:

Transaction (ObjID, Geometry, List((Key, [Value])))

By determining frequent itemsets from such transactions one obtains frequent patterns of attribute names (if key is the name of a property), tags (if key is a tagname) or words (if key is the word of a frequency term vector). These frequent patterns cannot yet be seen as concepts, but they are good candidates for building concept hierarchies and classification models in a subsequent step. The result of some frequent patterns in the OSM data, which can be interpreted as collaborative generated schemas for geographic concepts, is illustrated in Figure 1. The above process is discussed in more detail in (Sengstock and Gertz 2010).

3. Scale-Dependent Mining

A pattern is called frequent if it has a minimum support, that is, it occurs a minimum number of times in a given dataset. Patterns that only occur rarely are not considered. However, geospatial patterns occurring with a low frequency in a dataset can still be interesting for concept extraction. This is either because they are 1) densely clustered (and thus may represent a local/regional pattern on a large scale) or 2) they are widely distributed (and thus may represent a pattern on a small scale). We use a bottom-up approach based on a quadtree data structure to determine which items are candidates for itemset generation on a certain scale, as shown in Figure 2.



A	(coastline, type)
B	
C	(building,amenity)
F	(highway, name, housenumber)
G	(highway, name), (building, amenity)
J	(higway, name)

Figure 2. Scale-dependent frequent pattern mining.

The items of every transaction in each leaf node of the quadtree (which constitutes a grid of cells over the input data space) are counted. The items that occur in a cell with at least a minimum frequency are used to generate frequent itemsets over the transactions within this cell. On the next higher level all items that have not been used so far are summarized. If they reach the minimum frequency they are input to frequent itemset mining at this level. This step is repeated until the root node is reached. The determined itemsets are linked to the according nodes in the quadtree, which then also allows for fast exploration, for example, via a map interface.

4. Framework and OSM Explorer

Because of the complexity of frequent pattern mining algorithms it is necessary to pre-process the frequent itemsets to allow for efficient exploration. We developed a Frequent Pattern Store (see Figure 3) that handles the data import and transformation (Importer), the scale-dependent pattern mining (Frequent Pattern Miner) and the query processing of extracted patterns (Query Engine).

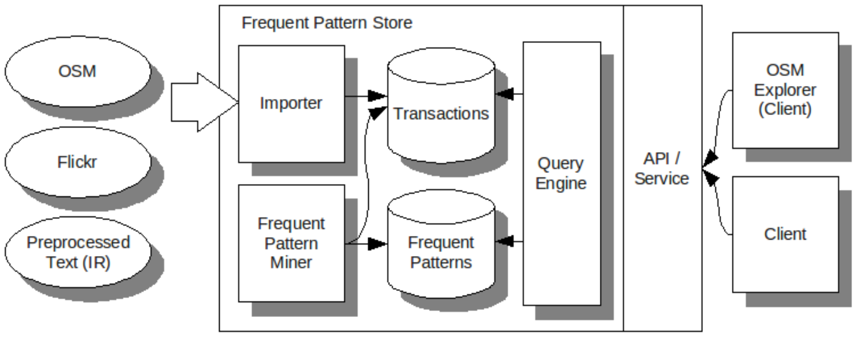


Figure 3. Frequent Pattern Store.

The framework can be deployed as a service and provides an API to allow ad-hoc queries for frequent patterns in a given region, and queries to retrieve the objects of a given frequent pattern. Editors can be extended to use the service for auto-completion of attribute names, calculation of quality measures, and the processing and visualization of available patterns. We implemented the OSM Explorer as a GUI with a map interface to visualize frequent patterns and quality measures (Figure 1). The GUI connects to the service and displays the frequent patterns of the map extent. The geoobjects of a selected frequent pattern are displayed on the map with additional information about quality measures and likely further attributes. For example, the consistency metric describes a measure for the distinctiveness of attributes in a given frequent pattern. A concept with high redundancy between the values of its attributes is assumed to be less consistent than a concept having attributes with clearly separated value domains.

5. Conclusions

We described a framework for exploring VGI based on scale-dependent frequent pattern mining. A motivation for our work is the fast and efficient integration of heterogeneous user-generated geodata and to merge all information available for certain geographic objects. Another motivation is to help users using VGI based on automatically generated quality measures and extracted concepts. A lot work needs to be done regarding the transformation of textual descriptions into transaction objects, and an evaluation of discovered patterns for several data sources needs to be conducted. Currently, we are building classification models on top of the determined patterns, to improve the extraction of concepts of the heterogeneous VGI.

References

Afrati F, Gionis A, Mannila H, 2004, Approximating a collection of frequent sets. Proceedings of KDD '04, 12-19

Deng D P, Chuang T R, Lemmens R, 2009, Conceptualization of Place via Spatial Clustering and Co-occurrence Analysis. Proceedings of the International Workshop on Location Based Social Networks, 49-55

Ding W, Eick C F et al, 2006, A Framework for Regional Association Rule Mining in Spatial Datasets. Proceedings of ICDM '06, 851-856

Edwardes A J and Purves S, 2007, A theoretical grounding for semantic descriptions of place. Proc. 7th Intern. Symp. on Web and Wireless Geographical Information Systems, LNCS 4857, 106-121

Gervais M, Bédard Y et al., 2009, Data Quality Issues and Geographic Knowledge Discovery. In: Miller H J, Han J (eds), Geographic Data Mining and Knowledge Discovery, 99-115

Goodchild M, 2007, Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211-221

Han J, Gao J, 2009, Research challenges for Data Mining in Science and Engineering. In: Kargupta H, Han J et al. (eds), Next Generation of Data Mining, 3-27

Koperski K and Han J, 1995, Discovery of Spatial Association Rules in Geographic Information Databases. Proc. 4th Intern. Symp. on Advances in Spatial Databases, LNCS 951, 47-66

Liu B, Hsu W, Ma Y, 1998, Integration of classification and association rule mining. Proc. KDD '98: 80-86

Sengstock C, Gertz M, 2010, Anwendung von Frequent Itemset Mining auf nutzergenerierte Geodaten. Geoinformatik 2010, Kiel, 28-36