universitätbonn

Clustering large spatial data sets

Dipl. Inf. Christian Pölitz

April 11, 2010

1 Task

Nowadays, data of positions where at certain times something happens becomes widely available. We call this kind of data, event data. This can be telephone calls, photos uploaded to Flickr or forest fires. The important aspect on the data is that it has spatial (mostly geographical) and temporal attributes.
Several data mining tasks can be applied to gain information out of these amounts of data. Pattern Mining methods can be used to find interesting movement patterns of the events. Similar structures among these events can be used for classification of new events. A very intuitive task for processing such data is to find dense regions. These regions can mean traffic jams where many drivers make calls on highways, interesting tourist places in case of photos from Flickr or endangered regions by forest fires. Due to the size of the data, traditional clustering methods will cost either large processing time or cannot be applied if all data does not fit into main memory.
The task of this lab group consists of implementation work and literature work. First, existing approaches of clustering large datasets as well as incremental clustering shall be studied. These existing methods must be analyzed according to how it can be applied to event data that we provide. Further, w.r.t. the possible solutions, the nature of the event data shall be used. This means possible improvements of the methods by exploiting the geographical and temporal attributes. For instance the geographical attributes can be used to build an index or we assume that the events are temporally ordered.
The students shall report their solutions for the task and give an additional presentation. This means the studied methods shall be summerized in a report, implemented in Java

in the stile of the Weka Library [1] and documented. The implementation must implement the Cluster interface from Weka for the clustering method. The data source shall be a flat file, but the implemention must be generatic enough to be easily integrate a database as source of the data. Eventually these results shall be sent to us and explained in a short presentation. The evaluation of the lab will be based on the implementation and the presentation.

2 Groups

We suggest that 2 or 3 students work together in a group. Each group must get familiar with the Weka libary and how to implement methods inside this library including the data structures used. We provide a short introduction at the first meeting. Further a group must solve one of the following clustering tasks:

- Incremental density based clustering: [EKS+98]
- An Improved Sampling-Based DBSCAN [EpKSX96] for Large Spatial Databases: [BB04]
- An Efficient Clustering Algorithm for 2D Multi-density Dataset in Large Database: [XWG07]

This means each group decides to solve one of these tasks by studying the referenced papers and implementing the described methods in the Weka library. For testing we provide a data set containing records with spatial and temporal attributes. In the results each data record shall be labeled with a cluster label and be written in a result file.

3 Contact and Credit Points

All students wishing to participate in this lab group write an email to:
poelitz@iai.uni-bonn.de
A successive accomplishment of this lab brings 8 Credit Points.

References

[BB04] B. Borah and D.K. Bhattacharyya. An improved sampling-based dbscan for large spatial databases. pages 92 – 96, 2004.

[EKS+98] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Michael Wimmer, and Xiaowei Xu. Incremental clustering for mining in a data warehousing environment. In *VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases*, pages 323–333, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[1] http://www.cs.waikato.ac.nz/ml/weka/

[EpKSX96] Martin Ester, Hans peter Kriegel, Jörg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[XWG07] Ying Xia, GuoYin Wang, and Song Gao. An efficient clustering algorithm for 2d multi-density dataset in large database. *Multimedia and Ubiquitous Engineering, International Conference on*, 0:78–82, 2007.