Event-based Clustering for Reducing Labeling Costs of Event-related Microposts

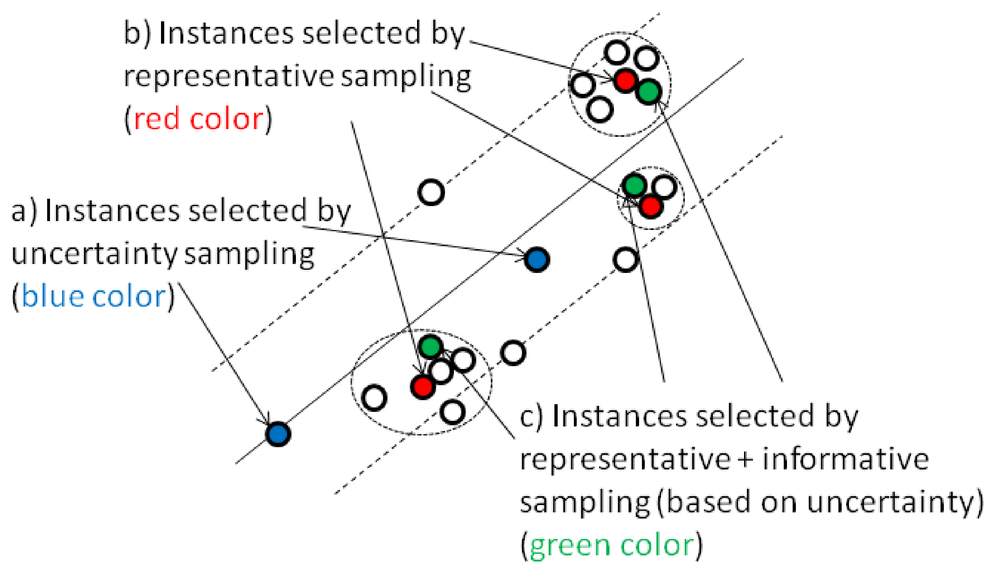
Axel Schulz[1,2], *Frederik Janssen*[4], Petar Ristoski[3], and Johannes Fürnkranz[4], [1]DB Mobility Logistics AG, Germany, [2]Telecooperation Lab, Technische Universität Darmstadt, [3]Data and Web Science Group, University of Mannheim, Germany, [4]Knowledge Engineering Group, Technische Universität Darmstadt

Motivation

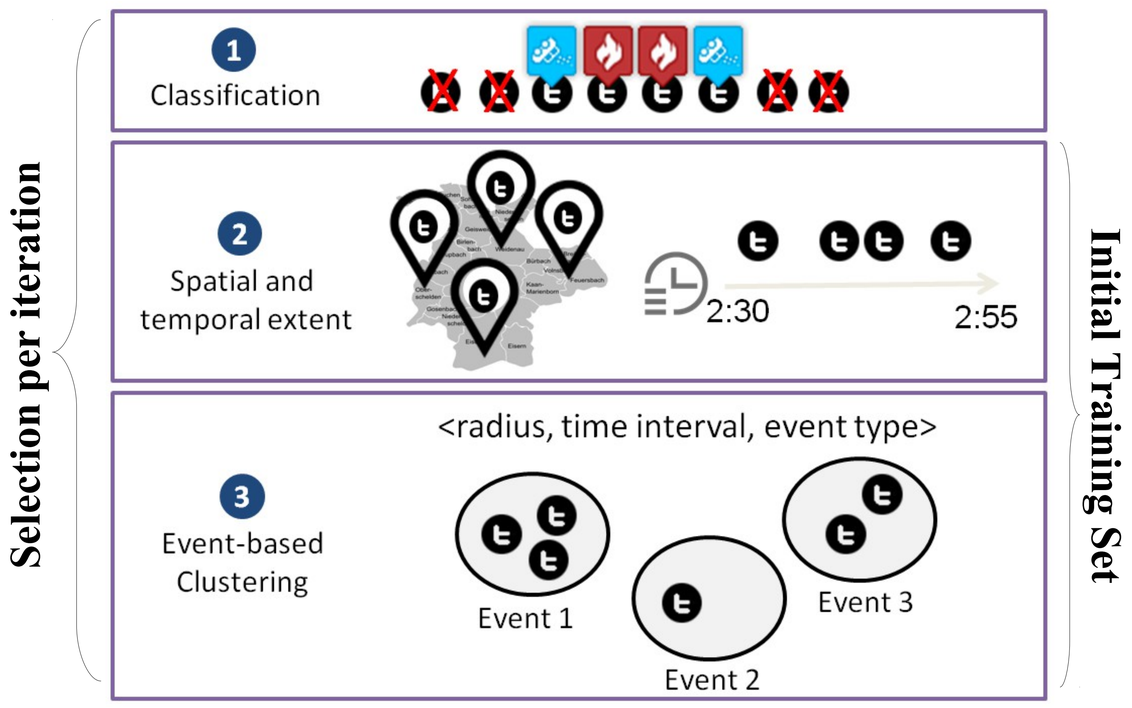
- event-related information very useful in domains like, e.g., emergency management
- main issue for supervised learning
  - obtaining labeled data very costly
- solution: **Active Learning**
  - needs *initial training set* and
  - method for *query selection per iteration*

Active Learning for Event Type Classification

- selection of most **informative** and **representative** instances
  - by using metadata for clustering



Event-based Clustering



Initial Training Set

- selection of informative instances not possible yet (***step 1***) → representative instances used
- apply Event-based Clustering based only on **spatial** & **temporal** extent
- order clusters by avg. *k*-nearest-neighbour-based density
- select instances from top to bottom
  - ensures selection of instances from best clusters, i.e., noisy clusters with unrelated items are avoided

Query Selection per iteration

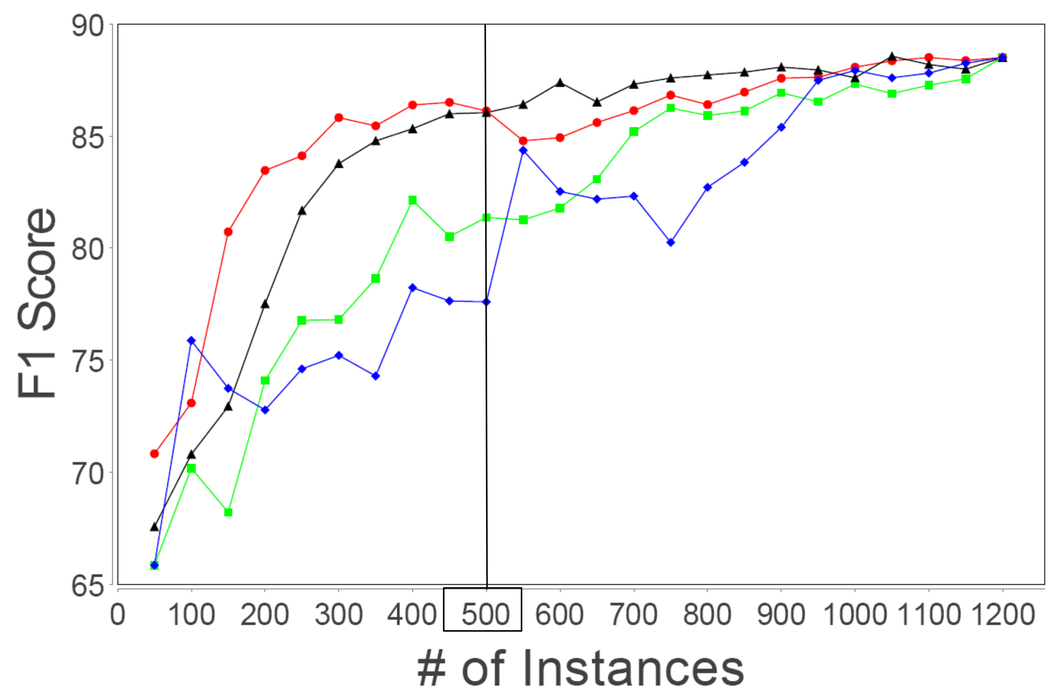
- train classifier on labeled data (***step 1***) and apply it on unlabeled data → assign **thematic dimension**
- apply Event-based Clustering based on all dimensions (***step 2*** *and* ***3***)
  - example rule: <Car_Crash, 200*m*, 20*min*>

- order clusters by avg. $DSH = \text{density} \times \text{entropy}$
- draw instances per cluster from top to bottom with logarithmic selection

Experimental Setup

- Event-based Clustering compared against three other approaches
  - Tang et al., 2002
    - *k*-means for initial clustering (k=4)
    - select most uncertain instances in each cluster
    - information density to weight examples
  - Zhu et al., 2008
    - *k*-means for initial clustering (k=4)
    - selection based on $\text{density} \times \text{entropy}$ measure
  - Uncertainty Sampling
    - random instances for initialization
    - selection strategy: entropy-based uncertainty sampling
- SVM classifier: weka's SMO
  - default parameters (same for all approaches)
- *classification accuracy* not main objective

Results



- better performance for initial selection (50 instances) & regions with few labeled instances (<500)

- drop after 500 instances: more instances result in higher # of clusters, rendering the selection more difficult

Approach	Deficiency
Tang et al., 2002	1
Zhu et al., 2008	0.90
Uncertainty Sampling	0.53
Event-based Clustering	**0.44**

- deficiency measures F1 of **all** iterations compared to baseline
- Event-based Clustering lowest value
- surprisingly good performance of Uncertainty Sampling
  - focusing only on informativeness good choice for this dataset

Conclusions and Future Perspectives

- **novel selection strategy** based on **temporal**, **spatial**, and **thematic** information
  - **better initial training set**
  - **improved selection** in each iteration
- future work: use framework in conjunction with labeling single features

References

- Min Tang, Xiaoqiang Luo, and Salim Roukos. Active learning for statistical natural language parsing. In ACL'02, pp. 120–127, 2002.
- Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In COLING'08, pp. 1137–1144, 2008.
- Axel Schulz. Mining User-Generated Content for Incidents. PhD thesis, TU Darmstadt, 2014. URL http: //tuprints.ulb.tu-darmstadt.de/4107/.