
**ERCIM "ALAIN BENSOUSSAN" FELLOWSHIP PROGRAMME**

ERCIM
European Research Consortium for Informatics and Mathematics

# Scientific Report

| | |
|---|---|
| First name / Family name | Max Zimmermann |
| Nationality | German |
| Name of the *Host Organisation* | SICS |
| First Name / family name of the *Scientific Coordinator* | Daniel Gillblad |
| Period of the fellowship | 01/06/2015 to 31/05/2016 |

## I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

### Online active learning for parallel computation:

Aiming to develop an online active learner for document stream classification we studied the problem of parallel computing in pool based active learning. We developed a concept that allows processing a batch of documents in parallel while distributing the two tasks (i) sampling new documents and (ii) predicting the label for them, across a set of workers. Selected documents are then used to update the classification model. The main contribution is the interplay between old documents (history) and arriving documents when selecting new documents to update the model.

### Privacy preserving for cellular network data:

We studied differential privacy, Kalman filter and Particle filter to anonymize a stream of cellular network data. In particular we proposed techniques that concentrate on the anonymization of sum queries such as the number of people travelling from A to B in a given time. The anonymized data is then used to apply analytic methods improving the prediction of traffic flow.

### Opinion stream mining (Active Learning):

We developed methods for opinion stream mining, i.e. document stream classification to predict the sentiment of documents. The focus was on active learning when the input is a stream of documents. We proposed the algorithm ACOSTREAM that adhered to the sequential strategy of active learning, i.e. the algorithm decides for each arriving instance whether it will request a label for it Our algorithm uses a variant of Multinomial Naive Bayes for classification, which deals with changes in the vocabulary of the arriving documents.

### Text classification for evolving streams:

We worked on a change resistant text classifier using an ensemble dirichlet compound multinomial model. We focused on changes of words in particular called virtual drift, i.e. the contribution/ frequency/ importance of words change according to the topic being discussed. We studied methods to update the class prior and the conditional likelihoods assuming a bayesian classifier. Furthermore, we used gradual forgetting stating that recent occurrences of words count more than historically old ones. We measured the frequency of a word by Exponentially
Weighted Moving Average (EWMA) and weighted the word accordingly. We ranked the impact of single words by their contribution to accurate class predictions.

### Scientific community contributions:

I was reviewer for the following journals:
- WIREs Data Mining and Knowledge Discovery
- IEEE Transactions on Knowledge and Data Engineering

## II – PUBLICATION(S) DURING YOUR FELLOWSHIP

### - Incremental Active Opinion Learning Over a Stream of Opinionated Documents; Max Zimmermann, Eirini Ntoutsi, Myra Spiliopoulou; Wisdom 2015

**Abstract:**
Applications that learn from opinionated documents, like tweets or product reviews, face two challenges. First, the opinionated documents constitute an evolving stream, where both the authors's attitude and the vocabulary itself may change. Second, labels of documents are scarce and labels of words are unreliable, because the sentiment of a word depends on the (unknown) context in the author's mind. Most of the research on mining over opinionated streams focuses on the first aspect of the problem, whereas for the second a continuous supply of labels from the stream is as- sumed. Such an assumption though is utopian as the stream is infinite and the labeling cost is prohibitive. To this end, we investigate the potential of active stream learning algo- rithms that ask for

labels on demand. Our proposed AC- OSTREAM 1 approach works with limited labels: it uses an initial seed of labeled documents, occasionally requests additional labels for documents from the human expert and incrementally adapts to the underlying stream while exploit- ing the available labeled documents. In its core, ACOSTR- EAM consists of a MNB classifier coupled with “sampling” strategies for requesting class labels for new unlabeled doc- uments. In the experiments, we evaluate the classifier per- formance over time by varying: (a) the class distribution of the opinionated stream, while assuming that the set of the words in the vocabulary is fixed but their polarities may change with the class distribution; and (b) the num- ber of unknown words arriving at each moment, while the class polarity may also change. 2 Our results show that ac- tive learning on a stream of opinionated documents, delivers good performance while requiring a small selection of labels.

- Opinion Stream Mining; Myra Spilopoulou, Eirini Ntoutsi, Max Zimmermann; Encyclopaedia of Machine Learning 2016

**Abstract:**
Opinion stream mining is a variant of stream mining, text mining and opin- ion mining. Its goal is learning and adaptation of a polarity model over a stream of opinionated documents. An “opinionated document” is a text as- sociated with a “polarity”. Polarity is a value that represents the “strength” and the “direction” of an opinion. The strength can be a categorical value (e.g. +, −) or a ranking value (e.g. zero to five stars) or a continuous value (e.g. in the interval [0, 1]). The direction refers to whether the opinion is positive, negative or neutral. Strength and direction are often mixed. For example, in a ranking using stars, five stars may stand for a very positive opinion, zero stars for a very negative one and three stars for a neutral one.

- Online active learning for concurrent computation; Max Zimmermann, Andrea Esuli; Expert System with Application 2016 (In Preparation)

- Change resistant text classifier based on ensemble DCM; Max Zimmermann, Daniel Gillblad; (In Preparation)

## III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

ICT TNG Postdoc Workshop (September 3rd, 2015)
SICS retreat 2015

# IV – RESEARCH EXCHANGE PROGRAMME (REP)

For the research exchange programme I selected the group of Fabrizio Sebastiani and Andrea Esuli at the CNR in Pisa/Italy. I stayed the week from 25th till 29th of 2016 at the Institute. The cooperation was very insightful. We were working on online active learning for concurrent computing. We are aiming a journal publication to summarize the output of my stay.