Stream Analysis of Incoming Events Using Different Data Analysis Methods

Matúš CIMERMAN*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xcimermanm@stuba.sk

Nowadays we often need to process and analyse massive amounts of data, popularly called Big Data. Big Data is often related to its three *V's*: *velocity, volume* and *variety* [1]. In last two decades there was a lot of research on parallel batch processing models and algorithms. This approach effectively deals with stationary massive data. When it comes streaming data, we need to think of stream as data evolving continuously and deal with all three characteristics of Big Data. Fourth, especially in streams, arrives as *variability of change* when our world changing rapidly in real time.

Lots of methods and research interests have been focused on knowledge discovery in stationary Big Data. We also name such methods *batch processing* suitable methods. However, today we are facing streaming data coming from different sources such as: *social media, sensors, network devices in large world networks* or *logging files.* Since these sources serving data in an unlimited manner and are potentially infinite, methods for batch processing often fail when it comes to stream processing [1].

In our work we focus on task *trend detection* in a data stream using different methods. We aim to make method *applicable* and *simple* to interpret for domain experts without having detail knowledge how model or used method works. When serving results and answers to domain expert, we focus to make visualization easy to interpret and understand.

Trend detection have concerned many analysts and mainly marketers to be able react and predict what is happening or will happen in future. Trends are typically driven by emerging events that attracts attention large fractions users [4]. Real-time trend detection is a crucial when we want to perform actions according to current trends fast. It makes sense detecting trends in real-time because its changes dynamically in time in natural manner.

* Supervisor: Jakub Ševcech, Institute of Informatics, Information Systems and Software Engineering

Spring 2016 PeWe Workshop, April 2, 2016, pp. 5–6.

Some methods specific for domain have been proposed and published for detecting trends, e.g. in social networks like Twitter. When detecting trend, we need take into account [2, 3]:

- *concept drift and change detection* as a process of identifying differences in the state of an object by observing it at different times. While in streaming context this presents process of segmenting data stream into different segments by identifying the points where stream dynamics change.
- *seasonal effects* can be short-term and long-term. Methods like *Holt-Winter* have been developed to deal with trend detection while consider seasonal component.
- *anomalies and spam* can be caused by data transmission errors or targeted violation, thus the these generally errors need to be identified because it can significantly affect data's value for real-time trend detection.
- *segmentation* and trend detection only for a chosen segment of users or categories.
- *forecasting and prediction* if there will be a trend or how will currently detected trend perform in future.

Our ultimate goal is to propose method which will provide semi-automatically selection of appropriate methods for trend detection in current domain. There are many techniques for detecting trends in batch manner. For example, methods like *Naive Bayes* or *decision trees* are not applicable as we know them in streaming data problems. We will try to appropriately alter selected well-known methods for streaming data problems. Core characteristic of such method are: *one-pass* on data, *real-time* answers and adjusting model, *horizontally scalable* and *in memory* to handle massive volume of data.

Finally, we propose to produce visualization for easy and fast understanding of detected trend. This visualization must be evaluated by user study for example using Eye-Tracking hardware. Evaluation of proposed method will be performed using synthetic dataset and source of real-time data.

References

[1] KREMPL, Georg, et al. Open challenges for data stream mining research. ACM SIGKDD Explorations Newsletter, 2014, 16.1: 1-10.

[2] Hill, D.J. & Minsker, B.S., 2010. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling and Software*, 25(9), pp. 1014-1022.

[3] Hulten, G., Spencer, L. & Domingos, P., 2001. Mining time-changing data streams. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, 18, pp. 97-106.

[4] Mathioudakis, M. & Koudas, N., 2010. TwitterMonitor : Trend Detection over the Twitter Stream. , pp. 1155-1157.