Data Mining Seminar Short Presentation Speech - Vaibhav Kasturia

Good afternoon everyone! I am Vaibhav, an ITIS student and am I here to talk about my topic "STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream"

Do you people know what is happening in the world right now? How would you find out events that are happening in real time. Mining the Twitter stream makes it possible to detect events before they are actually published as news article. Users also might be interested in either local news or global news, and either breaking news or news from the past week maybe. We need a mechanism that allows users to explore events across different time and space granularities. The algorithm designed for event identification should be a single pass algorithm, that is, the algorithm should avoid iterative computation and update the clustering results in an incremental way. Also, the clustering results must be human readable.

The authors of the paper to solve this problem, designed an Algorithm which they call STREAMCUBE, as it is an extension of the existing Data Cube Structure. It aggregates events in a spatio-temporal fashion and merges new events with the existing events in an incremental fashion. It is based on finding events based on hashtags in tweets, as hashtags in the tweet can more accurately be used to find out what the events is about rather than the words in the tweet. There is also the challenge of content evolving hashtags. For example, #merkel might be used today for describing some EU Conference event which she is attending but it might have been used in tweets in the past to describe her visit to the Hannover Messe. The feature of this algorithm is that it is able to handle such content evolving hashtags. Finally, it also is able to detect burst events and local events. Burst events are events which exhibit an unusually high popularity during a particular time period. An example of this could be the football world cup. Local events are events which are popular in a particular region.

Through the semester, I will be working on analyzing this algorithm in detail. Please feel free to ask any further questions.