Scalable Detection of Emerging Topics and Geo-spatial Events in Large Textual Streams

Erich Schubert, Michael Weiler, and Hans-Peter Kriegel

Institut für Informatik, LMU Munich, Germany
{schube,weiler,kriegel}@dbs.ifi.lmu.de

Social media are a popular source for live textual data. This data poses several challenges due to its size, velocity, and heterogeneity. Existing methods for emerging topic detection often are only able to detect events of a global magnitude such as natural disasters, or they can only monitor user-selected keywords or a curated set of hashtags. Interesting emerging topics may, however, be of much smaller magnitude and may involve the combination of two or more words that are not yet known in beforehand.

We present several contributions introduced in previous work [1, 2]:
(i) A significance measure that can detect emerging topics early, long before they evolve into "hot tags", by drawing upon experience from outlier detection.
(ii) An efficient online algorithm to track these statistics for all words and word-pairs with only a fixed amount of memory, and without predefined keywords.
(iii) The clustering of the detected co-trends into larger topics, because a single event will cause multiple word combinations to trend at the same time.
(iv) How to incorporate location information into this process to both allow reporting the locality of events as well as detecting local-only geo-textual patterns.

The significance score provides an estimated frequency and standard deviation of words, word-pairs, and word-location information on the data stream at minimal cost. It allows for normalization across location, culture, and language and enables the detection of change events both in already frequent and not previously seen combinations. In contrast to earlier work, it can monitor every word at every location with only a fixed amount of memory, compare the values to statistics from earlier data, and immediately report significant deviations with minimal delay. The algorithm is capable of reporting "Breaking News" in real-time as they happen in social media around the world. Location is modeled at different granularities, such that events can be detected at a city, country, or global level by incorporating OpenStreetMap data, or at particular coordinates.

References

[1] E. Schubert, M. Weiler, and H.-P. Kriegel. "SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds". In: *Proc. 20th ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*. 2014, pp. 871–880.

[2] E. Schubert, M. Weiler, and H.-P. Kriegel. "SPOTHOT: Scalable Detection of Geo-spatial Events in Large Textual Streams". In: *Proc. 28th Int. Conf. on Scientific and Statistical Database Management (SSDBM)*. 2016.