Noname manuscript No.
(will be inserted by the editor)

Temporal Social Media Analytics

Sihem Amer-Yahia (CNRS/LIG, France), Themis Palpanas (Univ. Paris Descartes, France), Mikalai Tsytsarau (Univ. of Trento, Italy), Sofia Kleisarchaki (LIG, France), Ahlame Douzal (LIG, France), Vassilis Christophides (Technicolor Labs, France)

Definition

Social Media represent a valuable source of subjective user-generated-content since they reflect opinions, beliefs, findings, or experiences of a large number of users on a wide range of topics. Temporal Analytics of social media content aims to provide insights regarding the dynamics of user conversations in different mining tasks over the vocabulary of words employed in the corresponding posts. For example, a time-aware analysis of social media posts will enable to recognize popular conversation *trends* over a period of time, to alert about *emerging topics* that are fast gathering momentum, to monitor *how topics* of particular interest *evolve*, to *trace changes* in key aspects of conversation summaries, such as *user opinions* and *sentiments*, or to *identify relationships* among these summaries (e.g., correlations).

Synonyms

Trend Detection, Novelty Detection, Change Detection, Concept Evolution

Historical Background

Temporal Analytics of social media can been seen as a branch of stream mining research for unstructured textual data [Aggarwal]. Several of the aforementioned social media analytics tasks can be formulated as variations of core data stream mining problems, such as *clustering*, *concept drift*, and *outlier detection*. For example an emerging topic is one that has not been observed before and reflects somewhat an anomaly in the input data stream. In this respect, we need effective and efficient clustering methods summarizing continuously the key characteristics of an incoming stream of posts. The formation of a new topic, seen as a cluster, corresponds to a new pattern in the underlying data featuring a novelty with respect to the history of the data stream. Eventually, this novelty may become a normal pattern, as more data points are added to this cluster. In other cases, the novelty may be isolated, and may not turn into a new pattern. To track topics evolution we actually need to monitor over time various distributions of words employed in users posts (frequency, co-occurrence, etc.) to detect changes that would trigger the creation of new clusters, the deletion of old ones and the splitting or merging of the retained.

A first bulk of research focused on stream extensions of state of the art clustering algorithms. To cope with the unbounded size and sequential data flow of streams the proposed techniques [cluStream, denStream, FlockStream] maintain in memory only the necessary statistics (snaphots) of the stream. Then to consider the evolving nature of data streams these techniques focus on the most recent significant changes - by applying fading (amnesic) functions and time windows of increasing granularity to discount past snapshots. An alternative approach considers evolution vectors where the changes of the stream are summarized and limited into detecting the creation of new clusters and the removal of old ones. Both techniques appear a weakness to segment the input stream into time breakpoints where significant changes arise. To this end, more sophisticated algorithms have been proposed solving the problem of change detection in social content by studying the statistical differences of clusters and data distribution inside consecutive and fixed-length time intervals [Hulten, kleisar].

Recently several algorithms have been proposed for continuous outlier detection [Sadik] where outliers are

Address(es) of author(s) should be given

reported at each time point among all the data points in the current sliding window. The reason is that an object may change its outlierness status during its lifetime. This entails that we need to continuously inspect each object that has not expired (either directly or indirectly) rather than inspecting it only once (e.g., when it arrives) [Yang, Angiulli, Kontaki]

Foundations

The first key challenge in temporal analytics is dynamic time segmentation in order to detect change (e.g., topic evolution, or change in aggregated opinion). The division of the input into equal fixed-length intervals induces a topic's evolution loss. For instance, a large interval is susceptible to miss a change as it absorbs the statistical changes inside its window, while a small interval may cause false alarms for frequently changing topics. Thus, the selection of a proper interval length is critical for effectiveness. Moreover, fixed-length intervals do not detect varying rates of change (e.g., hourly and daily changes), which might exist in evolving topics. Hence, the interval length needs to change dynamically as the topic evolves in order to capture changes in different time granularities (e.g., hour, day, week).

The second key challenge is to capture the variety of changes [BrzeziOski, ada] in social streams ranging from sudden to recurring, indicating the need for designing efficient and scalable methods and structures for mining the evolution.

The third key challenge is to provide scalable analysis methods that achieve a reasonable trade-off between time complexity and result accuracy in identifying change. In this case, performance can be measured in terms of the time proximity of the results to the true values, reflecting the delay with which the system detects some phenomenon.

Key Applications

In industry and politics, it is common, for people to follow what is being said in social media about a particular brand, a political candidate, or a campaign, based only on their bursty activity. Then tracking how particular topics of interest have evolved, possibly in response to marketing or PR interventions, is also frequently needed.

Various commercial social media monitoring tools rely on individual keywords, whose usage has rapidly increased in recent time, as proxies for emerging topics. Such a methodology has obvious limitations in characterizing the separate strands of conversations that may have simultaneously co-emerged in the data stream. For this reason several research proposals and systems have been proposed to address these challenges by considering different time and space tradeoffs to design a satisfactory service for the end-user.

Trend Detection: Trends are a core notion of Social Media analytics.[Benhardus] define a trend as a word or phrase that is experiencing an increase in usage, both in relation to its long-term usage and in relation to the usage of other words. Twitter maintains its own trending topic list which includes all-time popular topics of discussion.

[Saha] estimates trend components of topics using Hodrick-Prescott Trend Filtering and then figure out whether a topic is emerging by introducing a margin based loss function which penalizes static or decaying topcs. [Goorha] are looking for trending words that co-occur with a product of interest. According to them, an interesting phrase should both be mentioned frequently and should be relatively unique to the product with which it is associated. However, to compute the significance score of a phrase co-occurring with a product, the global phrase occurrence frequency is required, which is difficult to compute in streaming way. For this reason in [Naaman], introduce a method of trend discovery based on burst detection. The method is based on the observed frequency of a term in a current window (of fixed length), the mean and the standard deviation of the frequencies of the term over this window. Given these three values the trending score of the terms in the current window can be effectively computed in a streaming mode, because the mean and the standard deviation can be computed without iteration over the frequencies of a term in every window, but updated as a new window is finished. In TwitterMonitor [Mathioudakis] "bursty" keywords, i.e., suddenly appearing in tweets at an unusually high rate) are grouped into trends based on their co-occurrences. Furthermore, TwitterMonitor extracts additional information from the tweets that belong to the trend, aiming to discover interesting aspects of it, such as not, bursty keywords that however provide contextual information, frequently mentioned entities containing the trend keywords, links in related tweets, frequent geographical origins of related tweets. Compared to TwitterMonitor, EnBlogue [Alvanaki] additionally considers shifts in hashtag correlations. In particular, for each tag pair that contains at least one seed tag (selected by popularity), it tracks correlations. If the current correlation is significantly different from the prediction based on the previous correlation values a shift detection is reported (sudden but significant increases in the correlation of tag pairs).

User Opinion Evolution: [Varlamis et al.] propose clustering accuracy as an indicator of the Social Media topic convergence. Clustering accuracy measures the relative separation of the cluster centers with respect to cluster sizes and a number of unclustered entries (noise). By analyzing how accurate the clustering is in different time intervals, one can estimate how correlated or diverse user entries are.

The study of [Thelwall et al.] indicates that changes in Social Media sentiment are mainly caused by external events, resulting in synchronized, correlated or anti-correlated sentiments of various groups of people. Moreover, sometimes these changes are particularly small, making it necessary to apply more sophisticated methods capable of detecting such correlations under high noise conditions.

Recent studies have explored the problems of automated discovery of sentiment-based contradictions (i.e., topics and time intervals during which very diverse sentiments have been expressed on some given topic, or the general sentiment changes drastically) [Tsytsarau et al. DiversiWeb11], and of explaining these contradictions in terms of demographics [Tsytsarau et al. Sigmod14] and news events [Tsytsarau et al. KDD14].

Future Directions

Cross References

Recommended Reading

[cluStream] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In Proceedings of the 29th international confer- ence on Very large data bases - Volume 29, VLDB 03, pages 8192. VLDB Endowment, 2003.

[denStream] F. Cao, M. Ester, W. Qian, A. Zhou. Density-Based Clustering over an Evolving Data Stream with Noise

[FlockStream] Agostino Forestiero, Clara Pizzuti, Giandomenico Spezzano. A single pass algorithm for clustering evolving data streams based on swarm intelligence

[kleisar] S. Kleisarchaki, D. Kotzinos, I. Tsamardinos, and V. Christophides. A methodological framework for statistical analysis of social text streams. In Information Search, Integration and Personalization, LNCS, pages 101-110. Springer Berlin, 2013.

[Hulten] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In KDD01, San Francisco, CA, Aug. 2001.

[BrzeziOski] Dariusz BrzeziOski. Mining Data Streams With Concept Drift.

[ada] Iris Ada, Michael R. Berthold. EVE: a framework for event detection

[Statstream] Y. Zhu and D. Shasha, Statstream: statistical monitoring of thousands of data streams in real time, in VLDB, 2002, pp.358369.

[Zliobaite] Indre Zliobaite. Learning under Concept Drift: an Overview

[dong] Guozhu Dong, Jiawei Han, Laks V.S. Lakshmanan, Jian Pei, Haixun Wang, Philip S. Yu. Online Mining of Changes from Data Streams: Research Problems and Preliminary Results

[Hawwash] Basheer Hawwash, Olfa Nasraoui. StreamDashboard: A Framework for Mining, Tracking and Validating Clusters in a Data Stream

[Kifer] Daniel Kifer, Shai Ben-David, Johannes Gehrke. Detecting Change in Data Streams.

[Mustafa] Ahmad Mustafa, Ahsanul Haque, Latifur Khan, and Michael Baron. Evolving Stream Classification using Change Detection

[Aggarwal] C. Aggarwal. 2014. Mining text and social streams: a review. SIGKDD Explor. Newsl. 15, 2 (June 2014)

[Sadik] S. Sadik and L. Gruenwald. 2014. Research issues in outlier detection for data streams. SIGKDD Explor. Newsl. 15, 1 (March 2014), 33-40.

[Yang] D. Yang, E. Rundensteiner, and M. Ward. Neighbor-based pattern detection for windows over streaming data. In EDBT, pages 529540, 2009. [Angiulli] F. Angiulli and F. Fassetti. Distance-based outlier queries in data streams: the novel task and algorithms. Data Mining and Knowledge Discovery, 20(2):290324, 2010. [Kontaki] M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, and Y. Manolopoulos. Continuous monitoring of distance-based outliers over data streams. In ICDE, pages 135146, 2011.

[J. Benhardus. Streaming trend detection in twitter. National Science Foundation REU for Artificial Intelligence, NLP and IR, 2010.]

[A. Saha and V. Sindhwani. Learning Evolving and Emerging Topics in Social Media: A Dynamic NMF approach with Temporal Regularization. In Proceedings of the 5th International Conference on Web Search and Data Mining (WSDM), 2012 .]

[S. Goorha and L. Ungar. Discovery of significant emerging trends. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 57-64. ACM, 2010.]

[M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on twitter. Journal of the American Society for Information Science and Technology, 62(5):902918, 2011.]

[M. Mathioudakis and N. Koudas. TwitterMonitor: trend detection over the twitter stream. SIGMOD '10 http://www.l2f.inesc-id.pt/ fmmb/wiki/uploads/Work/misnis.ref11.pdf]

[F. Alvanaki, M. Sebastian, K. Ramamritham, and G. Weikum. EnBlogue: emergent topic detection in web 2.0 streams. SIGMOD '11]

[Varlamis et al.] Varlamis I, Vassalos V, Palaios A (2008) Monitoring the evolution of interests in the blogosphere. In ICDE Workshops, IEEE Computer Society, pp 513-518

[Choudhury et al.] examine sentiment biases in blogosphere's communities, relying on the entropy measure as an indicator of the diversity in opinions.

[Choudhury et al.] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann, Multi-scale characterization of social network dynamics in the blogosphere, in CIKM, 2008, pp. 1515-1516.

[Thelwall et al.] M. Thelwall, K. Buckley, and G. Paltoglou, Sentiment in twitter events, JASIST, vol. 62, no. 2, pp. 406-418, 2011.

[Tsytsarau et al. DiversiWeb11] Mikalai Tsytsarau, Themis Palpanas, Kerstin Denecke. Scalable Detection of Sentiment-Based Contradictions. International Workshop on Knowledge Diversity on the Web (DiversiWeb), in conjunction with the World Wide Web Conference (WWW), Hyberabad, India, March 2011

[Tsytsarau et al. Sigmod14] Mikalai Tsytsarau, Sihem Amer-Yahia, Themis Palpanas. Efficient Sentiment Correlation for Large-Scale Demographics. ACM SIG International conference on Management of Data / Principles of Database Systems (SIGMOD/PODS), New York, NY, USA, June 2013

[Tsytsarau et al. KDD14] Mikalai Tsytsarau, Themis Palpanas, Malu Castellanos. Dynamics of News Events and Social Media Reaction. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), New York, NY, USA, August 2014