MONIC and Followups on Modeling and Monitoring Cluster Transitions

Myra Spiliopoulou[1] **, Eirini Ntoutsi[2], Yannis Theodoridis[3], and Rene Schult[4]

[1] Otto-von-Guericke University of Magdeburg, Germany,
myra@iti.cs.uni-magdeburg.de,
[2] Ludwig-Maximilians-University of Munich, Germany (work done while with [3]),
ntoutsi@dbs.ifi.lmu.de
[3] Department of Informatics, University of Piraeus, Greece,
ytheod@unipi.gr
[4] Mercateo AG, Germany (work done while with [1]), rene.schult@mercateo.com

Abstract. There is much recent discussion on data streams and big data, which except of their volume and velocity are also characterized by volatility. Next to detecting change, it is also important to interpret it. Consider customer profiling as an example: Is a cluster corresponding to a group of customers simply disappearing or are its members migrating to other clusters? Does a new cluster reflect a new type of customers or does it rather consist of old customers whose preferences shift? To answer such questions, we have proposed the framework MONIC [20] for modeling and tracking cluster transitions. MONIC has been re-discovered some years after publication and is enjoying a large citation record from papers on community evolution, cluster evolution, change prediction and topic evolution.

Keywords: cluster monitoring, change detection, dynamic data, data streams, big data

1 Motivation

MONIC stands for MONItoring Clusters. It has appeared in [20] with following motivation:"In recent years, it has been recognized that the clusters discovered in many real applications are affected by changes in the underlying population. Much of the research in this area has focused on the *adaptation* of the clusters, so that they always reflect the current state of the population. Lately, research has expanded to encompass *tracing and understanding* of the changes themselves, as means of gaining valuable insights on the population and supporting strategic decisions. For example, consider a business analyst who studies customer profiles. Understanding how such profiles change over time would allow for a long-term proactive portfolio design instead of reactive portfolio adaptation."

** Work on model monitoring (2013) partially supported by the German Research Foundation, SP 572/11-1 *IMPRINT: Incremental Mining for Perennial Objects*.

Seven years later, this motivation holds unchanged and the need for evolution monitoring is exarcebated through the proliferation of big data. Next to "Volume" and "Velocity", "Volatility" is a core characteristic of big data. Data mining methods should not only adapt to change but also describe change.

Citations to MONIC (according to Google scholar) follow a rather unusual distribution. There has been one citation to the article in 2006 and 11 in 2007. From 2008 on though, 15-30 new citations are added every year, reaching 132 in 2012 and achieving 141 in 2013. This means that the article has been rediscovered in 2008, reaching a peak in popularity increase in 2011 (31 citations) and remaining stable thereafter. We associate these values with the intensive investigation of the role of time in data mining and the study of drift in stream mining. The topics associated with MONIC are multifaceted. MONIC is cited by papers on community evolution, on evolution of topics in text streams, on cluster evolution in general (for different families of cluster algorithms) and on frameworks for change modeling and change prediction. An important followup in terms of specifying a research agenda is the work of Boettcher et al [3] on "exploiting the power of time in data mining".

2 MONIC at a Glance

MONIC models and traces the transitions of clusters built upon an accumulating dataset. The data are clustered at discrete timepoints, cluster transition between consecutive time points are detected and "projected" upon the whole stream so as to draw conclusions regarding the evolution of the underlying population.

To build a monitoring framework for streaming data, the following challenges should be addressed: (a) what is a cluster? (b) how do we find the "same" cluster at a later timepoint? (c) what transitions can a cluster experience and (d) how to detect these transitions?

Regarding challenge (a), in MONIC a cluster is described by the set of points it contains. MONIC is thus not restricted to a specific cluster definition and can be used equally with partitioning, density-based and hierarchical clustering methods. More importantly, this cluster description allows for different forgetting strategies over a data stream, and even allows for changes in the feature space. Feature space evolution may occur in document streams, where new features/words may show up. Hence, MONIC is flexible enough for change monitoring on different clustering algorithms with various data forgetting models.

Challenge (b), i.e. tracing a cluster at a later timepoint is solved by computing the overlap between the original cluster and the clusters built at the later timepoint. The overlap is a set intersection, except of considering only those cluster members that exist at both timepoints. Hence, the impact of data decay on cluster evolution is accounted for.

Cluster evolution, challenge (c), refers to the transitions that may be observed in a cluster's "lifetime". The simplest transitions are disappearance of an existing cluster, and the emerging of a new one. MONIC proposes a typification of cluster transitions, distinguishing between *internal transitions* that affect the cluster itself and *external transitions* that concern the relationship of a cluster to other clusters. Cluster merging, the absorption of a cluster by another, the split of a cluster into more clusters, as well as cluster survival, appearance and disappearance are external transitions. An internal transition is a change in the cluster's properties, such as its cardinality (shrink, expand), compactness or location (repositioning of its center, change in the descriptors of its distribution).

Next to specifying a list of transitions, there is need for a mechanism detecting them. Challenge (d) is dealt through *transition indicators*. These are heuristics which may be tailored to a specific algorithm, e.g. by tracing movements of a cluster's centroid, or algorithm-independent, e.g. by concentrating only on the cardinality of the cluster and the similarity among its members. The transition indicators are incorporated in a transition detection algorithm that takes as input the clusterings between two consecutive timepoints and outputs the experienced cluster transitions. The algorithm first detects external transitions corresponding to "cluster population movements" between consecutive timepoints and for clusters that "survive" at a later timepoint, internal transitions are detected. Based on the detected transitions, temporal properties of clusters and clusterings, such as lifetime and stability, are derived and conclusions about the evolution of the underlying population are drawn.

The low complexity of the algorithm, $\mathcal{O}(K^2)$, where K is the maximum number of clusters in the compared clusterings, makes MONIC applicable to real big volatile data applications. The memory consumption is also low, only the clusters at the compared timepoints are required as input to the clustering algorithm, whereas any previous clustering results is removed.

In the original paper, we investigated evolution over a document collection, namely ACM publications on Database Applications (archive H2.8) from 1998 to 2004 and studied particularly the dominant themes in the most prominent subarea of database applications, namely "data mining". In [19, 14], we extended MONIC into MONIC +, which in contrast to MONIC that is a generic cluster transition modeling and detection framework, encompasses a typification of clusters and cluster-type-specific transition indicators, by exploiting cluster topology and cluster statistics for transition detection. Transition specifications and transition indicators form the basis for monitoring clusters over time. This includes detecting and studying their changes, summarizing them in an *evolution graph*, as done in [16, 15], or predicting change, as done in [17].

3 MONIC and Beyond

MONIC has been considered in a variety of different areas and applications. Rather than elaborating on each paper, we explain very briefly the role of evolution in each area, picking one or two papers as examples.

Evolution in social networks: There are two aspects of evolution in social networks [18], best contrasted by the approaches [18, 21, 4]. In [21] the problem of community evolution is investigated from the perspective of migration of individuals. They observe communities as stable formations and assume that an individuum is mostly observed with other members of the own community and rarely with members of other communities, i.e. movement from one community/cluster to another is rare. This can be contrasted to the concept of community evolution, as investigated in [4], where it is asserted that the individuals define the clusters, and hence the clusters may change as individuals migrate. The two aspects of evolution are complementary: the one aspect concentrates on the clusters as stable formations, the other on the clusters as the result of individuals' movements.

Frameworks for change prediction & stream mining: Due to the high volatile nature of modern data, several generic works for change detection or incorporation of time in the mining process have been proposed. In [7] the problem of temporal relationship between clusters is studied and the TRACDS framework is proposed which "adds" to the stream clustering algorithms the temporal ordering information in the form of a dynamically changing Marcov Chain. In [6] the problem of cluster tracing in high dimensional feature spaces is considered. Subspace clustering is applied instead of full dimensional clustering thus associating each cluster with a certain set of dimensions. In [17] the MEC framework for cluster transition detection and visualization of the monitoring process is proposed. Other citations in this area include [2, 16, 15].

Spatiotemporal data: With the wide spread usage of location aware devices and applications, analyzing movement data and detecting trends is an important task nowadays. The problem of convoy discovery in spatiotemporal data is studied in [10], a convoy being a group of objects that have traveled together for some time. The discovery of flock patterns is considered in [22], defined as all groups of trajectories that stay "together" for the duration of a given time interval. Continuous clustering of moving objects is studied in [9]. Other citations in this area include [1, 13].

Topic evolution: Topic monitoring is a necessity nowadays due to the continuous stream of published documents. In [11] a topic evolution graph is constructed by identifying topics as significant changes in the content evolution and connecting each topic with the previous topics that provided the context. [8] studies how topics in scientific literature evolve by using except for the words in the documents the impact of one document to the other in terms of citations. In [23] a method is proposed for detecting, tracking and updating large and small bursts in a stream of news. Recently [24], the method has been coupled with classifier counterparts for each topic in order to study dynamics of product features and their associated sentiment based on customer reviews. Other citations are [5, 12].

4 Tracing Evolution Today

Scholars become increasingly aware on the importance of understanding evolution. This is reflected in the increasing number of citations on MONIC and in the diversity of the areas it is cited from. Modeling evolution, summarizing it and predicting it are cornerstone subjects in learning from data. Big data, characterized by volatility, will contribute further to the trend.

References

1. H. H. Aung and K.-L. Tan. Discovery of evolving convoys. In *SSDBM*, pages 196–213, 2010.
2. A. Bifet, G. Holmes, B. Pfahringer, P. Kranen, H. Kremer, T. Jansen, and T. Seidl. Moa: Massive online analysis, a framework for stream classification and clustering. *Journal of Machine Learning Research - Proceedings Track*, 11:44–50, 2010.
3. M. Böttcher, F. Höppner, and M. Spiliopoulou. On exploiting the power of time in data mining. *SIGKDD Explorations*, 10(2):3–11, 2008.
4. T. Falkowski, J. Bartelheimer, and M. Spiliopoulou. Mining and visualizing the evolution of subgroups in social networks. In *Web Intelligence*, pages 52–58, 2006.
5. A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou. Topic evolution in a stream of documents. In *SDM*, 2009.
6. S. Günnemann, H. Kremer, C. Laufkötter, and T. Seidl. Tracing evolving clusters by subspace and value similarity. In *PAKDD*, pages 444–456, 2011.
7. M. Hahsler and M. H. Dunham. Temporal structure learning for clustering massive data streams in real-time. In *SDM*, pages 664–675, 2011.
8. Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles. Detecting topic evolution in scientific literature: how can citations help? In *CIKM*, pages 957–966, 2009.
9. C. S. Jensen, D. Lin, and B. C. Ooi. Continuous clustering of moving objects. *TKDE*, 19(9):1161–1174, 2007.
10. H. Jeung, M. L. Yiu, X. Zhou, C. S. Jensen, and H. T. Shen. Discovery of convoys in trajectory databases. *CoRR*, abs/1002.0963, 2010.
11. Y. Jo, J. E. Hopcroft, and C. Lagoze. The web of topics: discovering the topology of topic evolution in a corpus. In *WWW*, pages 257–266, 2011.
12. C. Lauschke and E. Ntoutsi. Monitoring user evolution in twitter. In *BASNA workshop, co-located with ASONAM*, 2012.
13. E. Ntoutsi, N. Mitsou, and G. Marketos. Traffic mining in a road-network: How does the traffic flow? *IJBIDM*, 3(1):82–98, 2008.
14. E. Ntoutsi, M. Spiliopoulou, and Y. Theodoridis. Tracing cluster transitions for different cluster types. *Control & Cybernetics Journal*, 38(1):239–260, 2009.
15. E. Ntoutsi, M. Spiliopoulou, and Y. Theodoridis. Summarizing cluster evolution in dynamic environments. In *ICCSA*, 2011.
16. E. Ntoutsi, M. Spiliopoulou, and Y. Theodoridis. FINGERPRINT summarizing cluster evolution in dynamic environments. *IJDWM*, 2012.
17. M. D. B. Oliveira and J. Gama. A framework to monitor clusters evolution applied to economy and finance problems. *Intell. Data Anal.*, 16(1):93–111, 2012.
18. M. Spiliopoulou. Evolution in social networks: A survey. In C. C. Aggarwal, editor, *Social Network Data Analytics*, pages 149–175. Springer, 2011.
19. M. Spiliopoulou, E. Ntoutsi, and Y. Theodoridis. Tracing cluster transitions for different cluster types. In *ADMKD workshop, co-located with ADBIS*, 2007.
20. M. Spiliopoulou, E. Ntoutsi, Y. Theodoridis, and R. Schult. MONIC – modeling and monitoring cluster transitions. In *KDD*, pages 706–711, 2006.
21. C. Tantipathananandh and T. Y. Berger-Wolf. Finding communities in dynamic social networks. In *ICDM*, pages 1236–1241, 2011.
22. M. R. Vieira, P. Bakalov, and V. J. Tsotras. On-line discovery of flock patterns in spatio-temporal data. In *GIS*, pages 286–295, 2009.
23. M. Zimmermann, E. Ntoutsi, Z. F. Siddiqui, M. Spiliopoulou, and H.-P. Kriegel. Discovering global and local bursts in a stream of news. In *SAC*, pages 807–812, 2012.

24. M. Zimmermann, E. Ntoutsi, and M. Spiliopoulou. Extracting opinionated (sub)features from a stream of product reviews. In *DS*, 2013.