EAI Endorsed Transactions
on Scalable Information Systems

Research Article ICST.ORG

Exploring Social Relationships in Text Streams

Ye Wang[1,2,*]

[1]Victoria University, Melbourne, Australia
[2]National University of Defense Technology, Changsha, China

Abstract

Mining social relationships offers us an opportunity to gain insights from non-obvious relationships between individuals. Its applications can be seen in various scenarios ranging from market planning, fraud detection to the protection of national security. Most raw information related to social relationships are continuously generated by social networks in a form of text, for the reason that it has the lowest storage consumption while still possesses powerful expression abilities. However, when these continuous texts are aggregated together forming enormous text streams, applying existing data mining approaches will encounter efficiency or usability issues, either due to their overlook of the dynamic property of streams or the inapplicability of traditional store-then-process paradigm. In this paper, we specify the research gap and present a review report for dynamic text streams.

Received on 29 July 2016; accepted on 29 July 2016; published on 05 August 2016

Keywords: Social Relationships, Text Streams, Social Network, Data Mining

doi:10.4108/eai.9-8-2016.151631

1. Introduction

Social network analysis, a research direction that investigates social structures through the use of network and graph theories has drawn significant attention from the academic community. Particularly, social relationships mining has been a hot issue because it reveals the implicit interpersonal relationships by studying complex interactions between individuals or groups. However, the advent of social media has greatly changed the landscape of mining process and brought up a series of new challenges.

First of all, since the analysis is extended from the traditional sociology context to the emerging data mining discipline, now anonymous and remote communications on the Internet are being considered in addition to the real connections in daily life, which makes the mining targets much more diverse and further enriches the implications of social relationships mining. Specifically, by examining the social network contents, we can identify a bunch of topics and events that are discussed by certain groups of users. The evolution of topics and events reveals the fact that communications and discussions between users can not only influence individuals, but also have a significant impact on the shifts of hot topics and relevant events over time. However, the internal mechanism that identifies the influences between individuals as well as the interaction between topics, events and individuals is relatively unexplored.

Secondly, the volume of source data has surged so greatly that the targeted data can no longer be stored somewhere before being processed. To enable social relationships mining on the continuously generated contents, we have to aggregate them together forming a data stream and process it in an incremental manner. Right now millions of users are recording their daily lives on the Internet, which results in enormous contents with various formats including texts, pictures and videos. However, the majority of substantive information is still distributed across the Internet using the plain text format due to its convenience and efficiency in expressing information. For this reason, our proposal has a dedicated mining data source of text streams.

Last but not the least, the timeliness of mining process has become much stricter than before. In some

★Please ensure that you use the most up to date class file, available from EAI at http://doc.eai.eu/publications/transactions/latex/

*Corresponding author. Email: ye.wang10@live.vu.edu.au

cases where the enormous contents in social networks are analysed for national security reasons, the value of hidden connections may vanish if warnings or further actions cannot be applied in time. Therefore, mining social relationships on social media has called for new mining theories and techniques that work on data stream and enable real-time relationships analysis.

Research challenges

In this proposal, we argue that topic, event and individual influences are the "three element of social relationships", based on which, we would like to address above demand by designing a new framework that can dynamically analyse the relationships between events and individuals over text streams. In particular, we firstly aim to study topic classification in social network text streams, through which different topics can be identified and tracked over time. Next, event evolution will be modelled and analysed in a view that resembles physical field, where public opinions are regarded as forces that have visible influences in social networks. Thirdly, we try to calculate individual influence within his/her social circle by taking the position of this individual in the whole diffusion networks into consideration. Finally, we plan to implement the real framework that contains the above mining techniques and is able to keep the real-time promise under possible fluctuated data streams. On the way towards achieving these aims, our research questions can be formulated as follows:

- *How to define, recognize and classify events and individual influences within dynamic and continuous data streams,*
- *How to quantitatively model the effect of evolving events having on individuals as well as its converse impact, and*
- *How to design novel data mining techniques that are capable of incrementally processing text streams with the above models and updating results in real time.*

At present, there are some existing works that have explored some certain parts of our research question. Topic Detection and Tracking (TDT) attempt to design methods that identify new and interesting events and follow the progress of previously identified events in long text corpus [1]. However, most of the existing works have not distinguished the two basic concepts of topic and event, resulting a certain extent of confusion [2]. Some other works try to evaluate the behaviors of network user by monitoring individual activities including posts, shares and profile update. Nevertheless, there are some important aspects of the social context that have been ignored, such as user's geographical location, sharing interval [3], and the consistency of attention from a group of users regarding to several different events.

In this paper, we specify the research gap and present a review report for dynamic text streams. Particularly, we identify the internal connections between elements of social network and employ the affiliation and derivation of those elements rather than ignoring or treating them sole as normally did previously.

The rest of the paper is organised as follows. Section 2 describes our contribution from academic and practical aspects. In section 3, we surveys related work in detail. Section 4 briefly proposes the intended methodologies and we finish with the conclusion in section 5.

2. Contributions

This section starts with discussing the academic contribution that this proposal intend to make to the existing mining theories. Then we explain certain practical benefits that are directly related to the social relationships mining and social network analysis.

2.1. Contribution to Knowledge

Our academic contribution on the meanings of social relationships mining is twofold. Firstly, the three elements of network interaction – topics, events, and individual influences will be clearly distinguished and identified from enormous social media contents, making it possible to investigate their internal connections and further guide possible predictions. Besides, though the related problems like topic detection and tracking, community detection and network structure analysis have been well studied from the perspective of individuals over the past decade; however, none of these studies discussed the problem of bridging existing research directions to reveal the inherent interactions. As a matter of fact, these mutual effects are becoming more significant than before, e.g. the public opinion now may has a great impact on the development of a specific event because of the increasing participation of the social media. Thus, it is important that research on social relationships cover these internal connections as an interactive whole rather than focusing on each separate side.

Secondly, we consider a lot of affiliated and derivative information in real-word social media scenarios when analysing the contents, such as the location and time-stamp of contents and the usage habits of individuals. Those information are normally neglected by the previous works because of lacking a formal model that can correlate these fragmented information in a meaningful way, plus that storing them consumes much more spaces compared to the main body of the posted text. In this proposal, we plan to contribute to filing in this gap by extending the processing paradigm of mining techniques from store-then-process to process-once-arrive, and designing new models that considers

affiliated and derivative information from the possible endless text streams.

2.2. Statement of Significance

Apart from theoretical contribution, our proposed research on social relationship mining will also have a strong practical significance.

To start with, our method differs from traditional social scientific studies in that we do not simply regard the attributes of individuals as the only determining factor in mining social relationships. On the contrary, we adopt an alternate view where the connection between individuals plays a much more important role than the attributes of individuals. The reason behind this choice is that, over the last decade, the popularity of social media has been greatly complicating the connections between individuals, resulting in a brand new propagation mode of information which is worth investigating to see how the public is being affected.

The practical significance of this proposal would be more intuitive if we put it into a real-world scenario. In the well-known revolutionary wave named "Arab Spring", social media contributed a lot in shaping political debates over the world, spreading awareness about ongoing events, and stimulating continuous online conversations on the process of political reform. The influence of social media was critical for both pro- and anti-governmental protesters to organize demonstrations, disseminate information of activities, and raise global awareness of their efforts. This is a perfect example on the bidirectional relationship between the cyber world and the real world. To be more specific, when the information of real activities got spread out by social medias, a large number of people will be actively or passively involved through their social relationships on the Internet. As time passes by, their participation, in turn, will have a significant impact on dominating the evolving trend of real world. In 2014, similar cases, "Umbrella Movement" and "Occupy Central", happened again in Hong Kong.

In this context, our work is indispensable because it is a necessity to be able to identify the relevant topics and events, as well as the recognition of particular individuals and groups that are involved in discussing and promoting the concerned real-world activities. The analysis gathering all these information together can shed lights on predicting how these activities will be evolved and how to propose appropriate countermeasures accordingly. It gives us the ability to stop crimes in their seed stage, for example, terrorists who exploit social network to devise and communicate their conspiracies can be identified and stopped before things getting worse. Also, it helps to reveal the certain group of people who are trying to manipulate the public opinion with ulterior motives, and restraining their activities is a crucial part of defending the national security.

3. Literature Review

Social network analysis (SNA) is theoretically rooted from early research on network analysis and social structure modeling which date back to 1930s. As its context continues to develop, now SNA has dealt with many aspects in various disciplines, including sociology, psychology and information science. The same pattern can also be seen in one of its sub-aspects – social relationships mining. However, though the scope of involved areas is being continuously expanded, the emergence of modern information technologies and the prevalence of online social modes has highlighted the three elements of social relationship in the mining process, calling for extensive research attention on each one of them.

Topic detection and tracking, as one of the representative outcomes, is used to reveal the clue of topics by examining a stream of broadcast news stories or a long text corpus [1, 4]. AlSumait et al. proposed an online topic model (On-Line LDA) [5] that uses a probabilistic approach to automatically identify emerging topics in text streams and further track their changes over time. Meanwhile, Makkonen et al. first presented the event evolution analysis as a sub-problem of TDT [6]. Most of the early research use the terms "topic" and "event" interchangeably which leads to an ambiguity to a large extent. Moreover, the earlier works do not take the user information into consideration and thus are unable to model the user-generated contents in online communities. As for events, the existing detection algorithms can be broadly classified into two categories: document-pivot based methods and feature-pivot based methods. The former detects events by calculating the semantics distance between documents [7], while the latter focuses on grouping the distribution of words to detect possible concealed events [8]. Yang et al. [9, 10] mined hierarchical event evolution graphs from news reports by considering three factors: content similarities, temporal features and document distributions, but their approaches heavily rely on the content-based analysis techniques which are not applicable to social media data.

Topic detection for online social media data requires the ability to deal with short, sparse and unformatted texts. There are several TDT methods developed to achieve that goal. For example, Sayyadi et al. [11] proposed an event detection algorithm to create a keyword graph for events discovery and description within the blogosphere. but the use of proposed algorithm requires a non-trivial tuning process of various parameters and thresholds, which induces usability barrier for potential users. In terms of

processing micro-blog posts, Cataldi et al. [12] leveraged a navigable topic graph that connects the emerging terms with other semantically related keywords to detect the emerging topics on Twitter, nevertheless, their definition of emerging keywords is controversial as only the words that frequently used in a given time period are considered. Saha and Sindhwani [13] proposed a framework that consists of online non-negative matrix factorizations to capture the evolution and emergence of themes on traditional TDT tasks. but in their work the online streaming data was only evaluated at the daily temporal scale. In the meantime, some other researchers placed an emphasis on identifying the individual threads to represent related events. Shahaf et al.[14] quantitatively formalised the concept of "coherence" among articles with linear structure and provided an algorithm to connect them. In their following work [15], a structure named "metro maps" was designed to concisely summarise the sets of documents which maximize the coverage of salient pieces of information. Though these methods are all suffered from inferior computational efficiency, they actually inspire us to study event evolution in highly interactive data. In addition, data representation should not only consider the textual features of the social messages, but also it is meant to capture the temporal, geographical, and social network information therein attached. For example, C. Aggarwal et al. [16] studied the problem of detecting and clustering events on social streams with two high-dimensional vectors, one of which denotes the textual word frequencies and the other hosts the recipients' social information. Recently, in the light of high interactivity and participatory of the social media data, Zafarani et al. [17] defined the sufficient and necessary conditions for emotion dissemination and first acknowledged the influence of public participation in the process of event detection. Deng et al. distinguished event information and public opinion in text corpus. They tracked the evolution of public concern which they believed is a reflection of corresponding events in social media data [18] and analysed the event evolution from the view of opinion field [19], however the authors failed to predict the future evolution of detected events.

Though the individual influence within the online social community has only been noticed for a few years, a broad spectrum of algorithms have already been proposed to meet the emerging requirements. Zafarani et al. [17] claimed that sentiment/emotion could propagate through a social network and subjective information posted by a user may affect the subjectivity of others. Weng et al. [20] proposed a topic-sensitive TwitterRank algorithm based on the well-known Pagerank algorithm [21] considering both the topical similarity between users as well as the link structures that connect user accounts. This algorithm is able to identify the topic-level influencers in the Twitter community. One of the major flaws that we found in the previous research is none of them has defined a formalised influence propagation model. Romero et al. [22] argued that an individual can not be influential unless he/her is able to overcome the user passivity. They also proposed an IP algorithm to estimate the influence and passivity of users based on their influence-passivity score. Taxidou [23] and Guille [24] analysed how information is diffused in social networks using the concept of information cascade. Another finding is that information cascade can help to identify the influential spreader by calculating the quantitative weight of individuals in the cascade.

As social media have greatly exacerbated the dynamism of our daily interaction, it becomes increasingly common that the data arriving at the analysis algorithm are possibly infinite continuous streams. The stream data can be delivered in two different forms, using either incremental data sequences [25, 26] or batches [27]. Though the former form would be closer to the reality of how data are generated in many real world applications, it also brings up more computational complexity and may encounter over-fit problem with test data sets. When taking the stability-plasticity dilemma [28] into consideration, most of the state-of-art works adopt the latter form [29–31] for the sake of stability. The term "stability" here means allowing the algorithm to retain any information that is still relevant, whereas the plasticity indicates that the relevant algorithm should be able to acquire new knowledge when needed.

"Topic classification" refers to the process of assigning text (or parts of the text) to a limited set of categories [32], which is the foundation of content-based social network analysis. The traditional methods for classification need to infer a static judgement function from the labeled training data set so that it can be further used to classify new data sets. In these models all the data that need to be classified is already stored at somewhere and ready to be accessed at any time. But as we have mentioned before, such store-then-process paradigm is apparently unrealistic for the emerging big data scenarios like online social network analysis, infectious disease control and stock exchange analysis. Due to the enormous size, source data collected in these scenarios has to be received as continuous streams, while applying traditional method on these streams will encounter a series of problems including concept drift and class imbalance.

Specifically, concept drift is a phenomenon that the classification boundaries may rapidly evolve along with the shift of public focus and not known a priori, which imposes a lot of difficulties in classifying streaming data. For example, before the disastrous earthquake

happened in Nepal on 25 April 2015, the news and posters related to "earthquake" were quite inactive compared to other trending keywords, however the number of news in this category soared right after the earthquake because it has immediately drawn the attention from all over the world. Soon afterwards, a similar emergency – the outbreak of MERS in Korea moved the focus of attention again from Nepal to east Asia and changed the boundaries of classification categories correspondingly. Hulten et al. [33] proposed the Concept-Adaptive Very fast Decision Tree (CVFDT), which keeps the classifier up-to-date by computing new split attributes and comparing them with the old attributes using sliding windows. SEA [34] trained the sub-classifiers on non-overlapping windows; however, the computational expense was still extraordinary under data of high evolving velocity. To achieve a trade-off between stability and plasticity, Wang et al. [35] assigned weights for sub-classifiers when dealing with new arriving data batches, but they refused to discard data that may still provide useful information. Elwell et al. [36] developed an Learn++.NSE algorithm to train a new classifier for each new batch of data, combining them using dynamically weighted majority voting. Nevertheless, the ability of identifying the ongoing concept drift of this kind of methods depends largely on the size of the windows or the batches and not specifically designed to accommodate class imbalance.

Class imbalance is another common challenge for classifying both static data and dynamic streams. There is a contrast in the volume of classes that some of them may have a plenty of entities, while some others are minor which has only limited number of elements. These minority classes are easy to be neglected to achieve high classification accuracy. However it would greatly impair the integrity of classification. For instance, in a case that the majority class accounts for 99.9% of the data set, a classifier that simply classifies all the elements into the majority class can achieve 99.9% accuracy. But by doing this, the minority classes are artificially ignored. As a matter of fact, these minority classes, such as the extremist views in public opinions, are at least as significant as the majority ones. In order to ensure the correctness of classification, they should not be sacrificed in exchange for accuracy [29]. To alleviate imbalance problem, one solution has been proposed to be working at the data-level which oversamples the minority instances [37] and heuristically undersamples the majority instances [38], while the rest solutions are mostly running at the algorithm-level, such as that Liu et al. improved the decision tree algorithm [39].

Though the problems of concept drift and class imbalance have gained extensive attention from the research community separately, the solution that efficiently solves them together to leverage the multilevel information is still long overdue. Ditzler and Polikar [31] proposed two ensemble approaches called Learn++.CDS and Learn++.NIE with different rebalance data strategies. They are proved to outperform earlier methods in terms of capability but at cost of much higher computational complexity. A learning framework for tackling the online imbalanced classification problem was proposed in [40] which consists of an imbalance detector, a concept drift detector and an online learner. It is then followed by a series of oversampling based online algorithms[41, 42] that further improved the performance. Recently, a selective re-train approach based on clustering was proposed [43] to re-train sub-classifiers when new data arrived. However, these existing methods either generate too many minority instances that overwhelm the majority, or induce excessive running overhead to accomplish their work.

4. Methodology

This section provides a discussion on the methodology that will be used to discover and connect the three elements of social relationships. We will also present an outlook of framework implementation built on top of real streams, where the possible defective stream data are properly handled on a scalable and fault-tolerance infrastructure.

The definitions of three elements are given below:

- A topic is a subject of conversation or discussion that happened on social network. It can be seen as a specific type of narrative, combing with all related materials and activities.
- An event is something that occurs in a certain place during a particular interval of time. The status of a single event can be categorized as either happened, ongoing or probable, which may change over time and got several distinct topics involved during its evolution.
- The individual influence is a quantization of the capacity to produce effects on the character, behavior, opinions, etc., of others in the same online community. Note that it can also evolve over time and affect the development of relevant events.

A new framework will consist the following three parts:

1. Multi-Window based Ensemble Learning for Topic Classification
2. Field View based Analysis and Prediction for Event Evolution
3. Evaluation Algorithm of Individual Influence.

5. Conclusions

Driven by the increasing demands of real-time relationships analysis in online text streams, we propose a framework that can dynamically identify and correlate the relationships between events and individuals. Specifically, we have modeled the "three element of social relationships" —topic, event and individual influence, based on which we explained how they connect with each other and further discussed the methodology that should be used to achieve our analysis targets.

References

[1] Allan J., Carbonell J., Doddington G. and Yang Y. (1998) Topic Detection and Tracking Pilot Study Final Report. *Proceedings of the Broadcast News Transcription and Understanding Workshop* (Sponsored by DARPA)

[2] Fiscus J. G. and Doddington G. R. (2002) Topic Detection and Tracking Evaluation Overview. In Allan J. [eds.] *Topic Detection and Tracking*, ser. The Information Retrieval Series (Springer US), 17–31.

[3] Kalyanam J. , Mantrach A., Saez-Trumper D., Vahabi H. and Lanckriet G. (2015) Leveraging Social Context for Modeling Topic Evolution. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD '15), 517–526. doi:10.1145/2783258.2783319.

[4] Makkonen J., Ahonen-Myka H. and Salmenkivi M. (2004) Simple Semantics in Topic Detection and Tracking. *Information Retrieval* **7**(3-4): 347-368. doi:10.1023/B:INRT.0000011210.12953.86.

[5] AlSumait L., Barbara D. and Domeniconi C. (2008) Online LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In *Proceeding of the 8th IEEE International Conference on Data Mining* (ICDM'08), 3–12. doi:10.1109/ICDM.2008.140.

[6] Makkonen J. (2003) Investigations on Event Evolution in TDT. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 Student Research Workshop - Volume 3*, (NAACLstudent '03), 43–48.doi:10.3115/1073416.1073424

[7] Yang Yiming, Pierce T. and Carbonell J. (1998) A Study of Retrospective and On-line Event Detection. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '98), 28–36.

[8] Kleinberg J. (2002) Bursty and Hierarchical Structure in Streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '02), 91–101.

[9] Yang C. C., Shi X. and Wei C.-P. (2006) Tracing the Event Evolution of Terror Attacks from On-Line News. *Intelligence and Security Informatics*, ser. Lecture Notes in Computer Science. (Springer Berlin Heidelberg), (3975): 343–354.

[10] Yang C.C., Shi X. and Wei C.-P. (2009) Discovering Event Evolution Graphs From News Corpora. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*. **39**(4): 850–863. doi:10.1109/TSMCA.2009.2015885.

[11] Sayyadi H., Hurst M. and Maykov A. (2009) Event Detection and Tracking in Social Streams. In *Proceedings of the Third International ICWSM Conference*.

[12] Cataldi Mario, Di Caro L. and Schifanella C. (2010) Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, (MDMKDD '10), 4:1–4:10.

[13] Saha A. and Sindhwani V. (2012) Learning Evolving and Emerging Topics in Social Media: A Dynamic Nmf Approach with Temporal Regularization. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, (WSDM '12), 693–702. doi:10.1145/2124295.2124376.

[14] Shahaf D. and Guestrin C. (2010) Connecting the Dots Between News Articles. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD '10), 623–632.

[15] Shahaf D. Guestrin C. and Horvitz E. (2012) Trains of thought: Generating information maps. In *Proceedings of the 21st international conference on World Wide Web*, 899–908.

[16] Aggarwal C. C. and Subbian K. (2012) Event Detection in Social Streams. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, **12**: 624–635. doi:10.1137/1.9781611972825.54.

[17] Zafarani R. Cole W. D. and Liu H. (2010) Sentiment Propagation in Social Networks: A Case Study in LiveJournal. *Advances in Social Computing*. ser. Lecture Notes in Computer Science. (Springer-Verlag Berlin Heidelberg), (6007): 413–420.

[18] Deng L., Xu B., Zhang L. Han Y., Zhou B. and Zou P. (2013) Tracking the evolution of public concerns in social media. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, 353–357.

[19] Deng L., Xu B., Zhang L. Han Y. and Zou P. (2013) Event Evolution Analysis in Microblogging Based on a View of Public Opinion Field. In *Proceeding of the Sixth International Symposium on Computational Intelligence and Design*, (ISCID), **2**:193–197. doi:10.1109/ISCID.2013.162.

[20] Weng J., Lim E.-P., Jiang J. and He Q. (2010) Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, 261–270.

[21] Page L., Brin S., Motwani R. and Winograd T. (2009) The PageRank citation ranking: bringing order to the Web. *Technical Report 1999-66, Stanford InfoLab*.

[22] Romero D. M., Galuba W., Asur S. and Huberman B. A. (2011) Influence and passivity in social media. *Machine learning and knowledge discovery in databases*. (Springer), 18–33.

[23] Guille A., Hacid H., Favre C. and Zighed D. A. (2013) Information diffusion in online social networks: A survey. *ACM SIGMOD Record*. **42**(2): 17–28.

[24] Taxidou I. and Fischer P. (2013) Realtime analysis of information diffusion in social media. In *Proceedings of the VLDB Endowment*, **6**(12): 1416–1421.

[25] Lange S. and Zilles S. (2003) Formal models of incremental learning and their analysis. In *Proceedings of the International Joint Conference on Neural Networks, 2003*, **4**: 2691–2696. doi:10.1109/IJCNN.2003.1223992.

[26] Fu L. (1996) Incremental knowledge acquisition in supervised learning networks. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*. **26**: 801–809. doi:10.1109/3468.541338.

[27] Gao J., Fan W., Han J. and Yu P. S. (2007) In *Proceedings of the 2007 SIAM International Conference on Data Mining*, doi:10.1137/1.9781611972771.1.

[28] Grossberg S. (1988) Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*. **1**: 17–61.

[29] Hoens T. R. and Chawla N. V. (2012) Learning in Non-stationary Environments with Class Imbalance. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD '12), 168–176. doi:10.1145/2339530.2339558.

[30] Hoens T. R., Polikar R. and Chawla N. V. (2012) Learning from streaming data with concept drift and imbalance: an overview. *Progress in Artificial Intelligence*. **1**(1): 89–101.

[31] Ditzler G. and Polikar R. (2013) Incremental Learning of Concept Drift from Streaming Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*. **25**(10): 2283–2301. doi:10.1109/TKDE.2012.136.

[32] Hillard D. and Purpura S. and Wilkerson J. (2008) Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics*. **4**(4): 31–46. doi:10.1080/19331680801975367.

[33] Hulten G., Spencer L. and Domingos P. (2001) Mining Time-changing Data Streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD '01), 97–106.

[34] Street W. N. and Kim Y. (2001) A Streaming Ensemble Algorithm (SEA) for Large-scale Classification. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD '01), 377–382. doi:10.1145/502512.502568.

[35] Wang H., Fan W., Yu P. S. and Han J. (2003) Mining Concept-drifting Data Streams Using Ensemble Classifiers. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD '03), 226–235.

[36] Elwell R. and Polikar R. (2011) Incremental Learning of Concept Drift in Nonstationary Environments. *IEEE Transactions on Neural Networks*. **22**(10): 1517–1531. doi:10.1109/TNN.2011.2160459.

[37] Chawla N. V., Bowyer K. W., Hall L. O. and Kegelmeyer W. P. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. **16**: 321–357.

[38] Tomek I. (1976) Two Modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics*. **SMC-6**(11): 769–772. doi:10.1109/TSMC.1976.4309452.

[39] Liu W., Chawla S., Cieslak D. A. and Chawla N. V. (2010) A robust decision tree algorithm for imbalanced data sets. In *Proceedings of SIAM International Conference on Data Mining*, 766–777.

[40] Wang S., Minku L. L. and Yao X. (2013) A learning framework for online class imbalance learning. *2013 IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL)*. 36–45. doi:10.1109/CIEL.2013.6613138.

[41] Wang S., Minku L. L. and Yao X. (2013) Online class imbalance learning and its applications in fault detection. *International Journal of Computational Intelligence and Applications*. **12**(4): 36-45 . doi:10.1142/S1469026813400014.

[42] Wang S., Minku L. L. and Yao X. (2015) Resampling-Based Ensemble Methods for Online Class Imbalance Learning. *IEEE Transactions on Knowledge and Data Engineering*. **27**(5): 1356–1368. doi:10.1109/TKDE.2014.2345380.

[43] Zhang D., Shen H., Hui T., Li Y., Wu J. and Sang Y. (2014) A Selectively Re-train Approach Based on Clustering to Classify Concept-Drifting Data Streams with Skewed Distribution. *Advances in Knowledge Discovery and Data Mining*. 413–424.