ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse47862019.pdf
https://doi.org/10.30534/ijatcse/2019/47862019

Towards an outlier detection model in text data stream

Awab Noori [1,4], Siti Sakira Binti Kamaruddin [2], Farzana Binti Kabir Ahmad [3]
[1] Data Science Research Lab, School of Computing, Universiti Utara Malaysia, Malaysia, a_noori_abdulwadood@ahsgs.uum.edu.my
[2] School of Computing, Universiti Utara Malaysia, Malaysia, sakira@uum.edu.my
[3] School of Computing, Universiti Utara Malaysia, Malaysia, farzana58@uum.edu.my
[4] Computer Technical Engineering, Alkitab University, Iraq.

ABSTRACT

This study proposes an outlier detection model in text data stream. Text stream is an important variant of data stream clustering. It has many useful implementations such as trend analysis, detection and tracking of topics, recommendation of user, and outlier detection. Outlier detection detects events which are interesting to the user and perhaps can be used to trigger some actions. One challenge in outlier detection in text stream is that normal behavior can change and thus it should be possible to adapt the models to the changes. Therefore, detecting outlier in text stream is not a trivial task. This paper proposes a conceptual model to detect outliers in the text stream. The model contains four main phases namely pre-processing, text representation, feature selection, and outlier detection phase. In the first phase, tokenization, stop words removal and stemming will be used. An incremental term weighting for text stream representation will be proposed in the second phase. An online feature selection will be improved on phase number three. Finally, in the fourth phase, one of the swarm intelligence techniques will be improved to detect outlier in the text stream.

Key words: Text stream clustering, swarm intelligence, outlier detection, online feature selection.

1. INTRODUCTION

Throughout its generation and diffusion, text became a primary medium for sharing information among the individuals. News content articles, blog/micro-blog posts, web-sites, educational journals, search engine queries/results, electronic mail, and computer records, are all examples of textual content information resources that are used in our daily interactions. A typical attribute among this kind of data is that they can be noticed as a text stream. Therefore, the text stream analysis became an essential issue with regard to the research of big data, because of its large volume [1]. The ability to cope with time frame has become an essential demand for the analysis of this particular continuous stream of the text because the available text data are not only huge and wealthy but also dynamic in nature [2]. For this reason, an incremental way is thus needed on textual data streams, which represent analysis techniques that are capable of working in such an environment. For instance, in discovering new topics as long as they arise from newly arriving text stream and within a specific time period. Because of its beneficial practical value, text stream analysis has gained great attention. The normal way to discover such huge unlabeled datasets is cluster analysis which is very helpful. Clustering text streams has several useful applications such as discovering spam emails, detecting new and tracking current news stories, user queries and discovering emergent trends in Twitter, document organization and topic detection and tracking, text crawling, and news group filtering. Accordingly, clustering text streams is a very significant part of the process of data mining. In addition to the text-based applications mentioned above, cluster analysis can be applied to detect outliers [1]. An outlier detection model is suggested in the text data stream in this paper. The purpose of outlier detection is to acknowledge novel or unfamiliar ideas from ongoing text streams. For example, the emergence of fresh hot subjects and the removal in a specific time frame of some other subjects relates to outlier detection in micro blogging web sites. Therefore, the most challenging issues in the area of data stream mining is outlier detection. As a result, outlier detection becomes an interesting subject for numerous researchers in the machine learning field. The next section details out the conceptual model proposed.

2. CONCEPTUAL MODEL

The conceptual model of the proposed outlier detection process in text data stream is illustrated in Figure 1. The model contains four main phases. The first phase is text pre-processing whereas representation of text is the second phase in the model. The feature selection is the third phase and the outlier detection phase is the final phase. In addition to the challenges mentioned above, text streams has many challenges such as dimensionality, unstructured, evolving text, variety, a high level of noise, concept drift, memory and processing time [3]. However, the biggest challenge in text stream mining is dimensionality.

Therefore, dimensionality reduction methods should be employed. Nevertheless, data dimensionality has a long road

ahead to tackle data streams context. At the same time, the literature of data streams show less work in this area. The majority of techniques actually developed to handle limited datasets and they are basically incremental algorithms.

Subsequently, the static reduction methods cannot be adapted to work with the data stream because the majority of these methods assume that all data are available, and the characteristics of the data will not change over time.

3. TEXT PRE-PROCESSING

In the knowledge discovery process, data preprocessing is one of the most significant steps. In fact, information pre-processing very often needs much more effort in the entire process of data preprocessing, more than the half effort. Data Pre-processing is the preparation method for the next stage of data processing. The most common steps in text preprocessing are tokenization, stop words removal and stemming. Tokenization step divide text stream to be words or phrases to construct informative elements called tokens. Whereas stopword removal is about removing the unusual words like pronouns and prepositions. By removing those words, the sequence of tokens will be decreased, and performance will be improved. Stemming means that lowering the word to its stems or root. Because some words have the same meaning but with different form. The following section describes the next phase of the proposed model.

4. TEXT REPRESENTATION

The next important step is term (feature) weighting for the purpose of text representation. Term weighting depends on some statistical information of terms that are used in a stream of text. One of the main issues in data stream analysis is making use of traditional TF-IDF weighting. As consequence, inappropriate labelling of novel or repetitive terms on a document may take place. Nevertheless, term use statistics may be unidentified, partial, or subject to drift in the streaming situation [4]. As a result, incremental TF-IDF [5] is used , due to the continuous updating of word utilization statistics, a sufficiently large set of documents proves to be effective. An online approach is presented in [5] to address the problem in data streams. However, they do not focus on text data. Therefore, as a solution to the issue mentioned above, an incremental term weighting scheme for text stream data will be proposed.

5. FEATURE SELECTION

Although existing feature selection techniques can be easily adapted to online scenarios, In the sense of high-dimensional mining data, the search space is dramatically increasing in size, resulting in an intractable computational requirement to obtain an optimal subset of features. Concept evolution and drifting feature space are other issues that traditional feature selection faces [6]. Additionally, feature selection algorithms such as [7] applied in the conventional text clustering reveals that static features are not appropriate for the text stream situation in the long-time condition.

Therefore, as a solution for this issue, in this research an online feature selection method will be improved. Contrary to offline counterparts, online filters do not require to ingest all data at once and appear to adapt well to drifts [6]. On top of that, online methods normally deal with issues derived from streams that simply the offline methods cannot solve it. Offline methods encounter difficulties derived from streams that could not handle it, which is not the case in online methods such as the arrival of new classes or features [8].

6.OUTLIER DETECTION PHASE

Outlier detection in the field of data stream is considered one of the most difficult issues [9]. Therefore, in the machine learning community, it has drawn the attention of many scientists [10]. The problem of outlier detection is closely related to text streams clustering. Text stream clustering algorithms can be used for outlier detection [11]. An online data analysis has emerged as a new form to analyse a data stream and detect outliers by using a clustering algorithm, where data changes over time. Therefore, density-based methods used for outlier detection in two ways. The first one, as independent points that do not fit into any of the clusters. The second one, clusters that considered as smaller than other clusters. In other words, in density-based methods, the small cluster that contains small data points and data points which are far from cluster centroid are considered as an outlier [12].

There are some approaches for text stream clustering in the literature attempting to produce much better outcomes from the current ones. However, most clustering approaches of the data stream use a two-phase approach to perform the clustering task. The two-phase clustering method, however, is computationally expensive, and repeated execution of the offline phase is needed to detect changes in the flow steam of data [13].

Furthermore, many text stream methods do not change their usability during the long-term clustering phase in the continuous variation of apps. Many methods of stream clustering have been applied to other kinds of data like categorical and text data [14]. Future work should therefore concentrate on researching certain types of datasets such as text streams [14]. While clustering has drawn much attention from research, there is little work on clustering information from text streams [15], [16]. For this purpose, new algorithms must be proposed or the current algorithms changed [17].

To deal with the limitations that exist in text stream clustering methods, swarm intelligence techniques (SI) seems as a promising avenue to tackle the above issues [13]. SI algorithm is a clustering algorithm class that mimics how nature works, inspired by biological examples such as bird flocks, bees, and ants. Such techniques should be tested for robustness and versatility in text stream mining. This particular classification has been quite favoured recently since it is possible to solve many problems without having strict mathematical methods. The motivation behind the use of SI to solve text stream clustering problem is due to its success found in addressing data stream clustering [13] and anomaly detection in data streams [15].

Outlier detection step is the last phase in the proposed model. In this phase an improved SI algorithm that will be able to detect outlier in text stream data is proposed. An improved SI algorithm is proposed in this step that will be able to detect outsiders in text stream. Every data object is associated with the agent in the improved algorithm. The agents are distributed randomly into a virtual space to move around independently to form one or more clusters. The outliers are identified in two different ways, either by the data point associated with agents grouped in a cluster of entities below a given threshold, or by data associated with isolated agents.

7.CONCLUSION

This study introduced a conceptual model to detect outlier in the text stream. The aim of this study is to improve the outlier detection rate in text stream data. This will be done by first performing the pre-processing steps such as tokenization, stop words removal and stemming. Then, utilizing an incremental term weighting to represent the text stream. After that, an online feature selection method is proposed to be used as feature reduction technique.

Finally, the swarm intelligence method will be enhanced to detect outlier in the textual data stream. The outcome of this study will contribute in revealing the extraordinary patterns in the textual data stream which will lead to better decision making. The results are expected to be promising when the proposed model is implemented on a real-life text stream data. Future work is going to concentrate on the implementation of the proposed conceptual model to make them more accurate to detecting outliers in text stream data.

REFERENCES

[1] B. Manoj, K. V. K. Sasikanth, M. V. Subbarao, and V. Jyothi Prakash, “Analysis of data science with the use of big data,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 6, pp. 87–90, 2018. https://doi.org/10.30534/ijatcse/2018/02762018

[2] O. Oueslati, A. I. S. Khalil, and H. Ounelli, “Sentiment analysis for helpful reviews prediction,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 3, pp. 34–40, 2018. https://doi.org/10.30534/ijatcse/2018/02732018

[3] A. Jain and I. Sharma, “Clustering of Text Streams via Facility Location and Spherical K-means,” in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 1209–1213. https://doi.org/10.1109/ICECA.2018.8474757

[4] X. Fu, E. Ch’ng, U. Aickelin, and L. Zhang, “An improved system for sentence-level novelty detection in textual streams,” 2015. https://doi.org/10.2139/ssrn.2828008

[5] C.-H. Lee, C.-H. Wu, and T.-F. Chien, “BursT: a dynamic term weighting scheme for mining microblogging messages,” in *International Symposium on Neural Networks*, 2011, pp. 548–557. https://doi.org/10.1007/978-3-642-21111-9_62

[6] S. Fong, R. Wong, and A. V Vasilakos, “Accelerated PSO swarm search feature selection for data stream mining big data,” *IEEE Trans. Serv. Comput.*, vol. 9, no. 1, pp. 33–45, 2016. https://doi.org/10.1109/TSC.2015.2439695

[7] A. Y. Al-Qammaz, F. K. Ahmad, and Y. Yusof, “Optimization of Least Squares Support Vector Machine Technique Using Genetic Multi-Dimensional Signals,” *J. Teknol.*, vol. 10, pp. 107–115, 2016.

[8] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, “A survey on data preprocessing for data stream mining: current status and future directions,” *Neurocomputing*, vol. 239, pp. 39–57, 2017.

[9] S. S. Kamaruddin, A. R. Hamdan, A. A. Bakar, and F. Mat Nor, “Deviation detection in text using conceptual graph interchange format and error tolerance dissimilarity function,” *Intell. Data Anal.*, vol. 16, no. 3, pp. 487–511, 2012. https://doi.org/10.3233/IDA-2012-0535

[10] E. R. Faria, I. J. C. R. Gonçalves, A. C. P. L. F. de Carvalho, and J. Gama, “Novelty detection in data streams,” *Artif. Intell. Rev.*, vol. 45, no. 2, pp. 235–269, Feb. 2016.

[11] C. C. Aggarwal, *Machine learning for text*. Springer, 2018. https://doi.org/10.1007/978-3-319-73531-3

[12] D. Haidar and M. M. Gaber, *Data Stream Clustering for Real-Time Anomaly Detection : An Application to Data Stream Clustering for Real-time Anomaly Detection : An Application to Insider Threats*, no. May. 2018. https://doi.org/10.1007/978-3-319-97864-2_6

[13] C. Fahy, S. Yang, and M. Gongora, “Finding multi-density clusters in non-stationary data streams using an ant colony with adaptive parameters,” in *Evolutionary Computation (CEC), 2017 IEEE*

Congress on, 2017, pp. 673–680. https://doi.org/10.1109/CEC.2017.7969375

[14] N. Masmoudi, H. Azzag, M. Lebbah, C. Bertelle, and M. Ben Jemaa, "CL-AntInc Algorithm for Clustering Binary Data Streams Using the Ants Behavior," *Procedia Comput. Sci.*, vol. 96, pp. 187–196, 2016.

[15] A. Forestiero, "Bio-inspired algorithm for outliers detection," *Multimed. Tools Appl.*, vol. 76, no. 24, pp. 25659–25677, 2017. https://doi.org/10.1007/s11042-017-4443-1

[16] P. Antonellis, C. Makris, Y. Plegas, and N. Tsirakis, "Efficient Algorithms for Clustering Data and Text Streams," in *Encyclopedia of Information Science and Technology, Third Edition*, IGI Global, 2015, pp. 1767–1776. https://doi.org/10.4018/978-1-4666-5888-2.ch170

[17] A. S. Hashmi, M. N. Doja, and T. Ahmad, "AN OPTIMIZED DENSITY-BASED ALGORITHM FOR ANOMALY DETECTION IN HIGH DIMENSIONAL DATASETS," vol. 19, no. 1, pp. 69–77, 2018. https://doi.org/10.12694/scpe.v19i1.1394