Data-Driven Behavioural Analytics in Web Logs Using Machine Learning Techniques

OPEN ACCESS

Volume: 13

Special Issue: 3

Month: February

Year: 2026

P-ISSN: 2321-788X

E-ISSN: 2582-0397

Citation:
S.Sathya, et al. "Data-Driven Behavioural Analytics in Web Logs Using Machine Learning Techniques." *Shanlax International Journal of Arts, Science and Humanities,* vol. 13, no. 3, 2026, pp. 430–39.

DOI:
https://doiorg/10.34293/sijash.v13iS3-i2-Feb.10316

Mrs. S. Sathya
Reasearch Scholar, Department of Computer Science
Alagappa University, Karaikudi

Dr. E. Ramaraj
Formaly Head and Professor, Department of Computer Science
Alagappa University, Karaikudi

Dr. V. Devi
Associate Professor, PG and Research Department of Computer Science
Thiruthangal Nadar college, Chennai

S. JayaSutha
Head and Assistant Professor
Department of Criminology and Criminal Justice Science
Thiruthangal Nadar college, Chennai

Abstract

Behaviour analysis of users on the internet is a crucial area that enables various features can be studied via user behaviour on the internet. The intention prediction is been recent research that identifies the user interactions on a website. Additionally, addressing the demand and enabling the information prediction for users enforces analysis of the user navigation behaviour. In this paper, we study the user behaviour in e-marketing websites to increase the relevance of bringing the products based on the user behaviour. The study uses a machine-learning algorithm with several metrics that studies the logs of several users during the training phase and provides user-specific relevant information during the testing phase. The simulation is conducted to test the efficacy of machine learning in providing the results relevant to the user behaviour, where accuracy is the main metric that is tested to address the machine learning ability. From the results, it is found that the proposed machine learning model achieves a higher rate of accuracy than other existing methods

Behaviour Analysis, Interactions, E-marketing, User Behaviour, Machine Learning

Introduction

The impact of the internet may be felt in virtually every sphere of human activity, including but not limited to education, commerce, and the retail sector, as well as all points in between. This is accomplished by collecting every mouse click and input from the user [1]. The use of proper marketing tactics, such as the pursuit of data such as user subscriptions, websites viewed, and so on, can help businesses increase their sales. This is because SEO strategies can

Pre-Processing; Pattern Discovery and Pattern Analysis as in Figure 1.

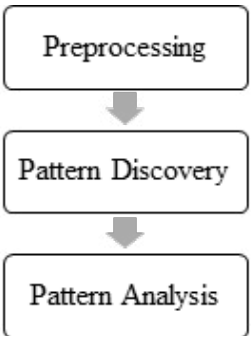

Fig. 1 Stages of Web Mining

Keeping a record of all server operations typically requires the use of a log file. The server keeps a record of this information in a variety of log files, including access logs, error logs, piping logs, script logs, and so on. These logs are organised into a hierarchical structure [8]. The access logs on the server are a record of all of the requests that the server has received and processed. Errors that take place in the course of processing requests are logged in error logs, which are consulted in the event of system malfunctions [9]. Piped Logs allow the server to write access and error logs directly to a process that is currently running, rather than publishing them to a file. Script Logs, which maintain an account of everything that goes into and comes out of the script, make it easier to debug and test CGI programmes [10]. Script logs also keep track of everything that comes out of the script. When we begin the process of mining online usage, the first thing we do is check the server access logs to see how people are navigating to the various web pages and services available on the internet [11].

This research aims to better understand how consumers interact with e-marketing websites in order to better target their purchases. The purpose of this research is to better understand how consumers interact with e-marketing websites. An algorithm for machine learning is utilised in order to perform the tasks of analysing log data from a large number of users during the training phase and giving user-specific relevant information during the testing phase. It is a test to determine whether or not machine learning can produce findings that are relevant to user behaviour, and the major criterion used to evaluate that capability is accuracy.

Related works

In the field of electronic commerce, the majority of data mining techniques rely on server logs to extract the sequences of user navigation events. On the other hand, these sequences are transformed into session characterizations as opposed to being extracted. It is usual practise for the characterization to include a set of high-level statistics that characterise the trip taken by the user. This kind of structure can accommodate a diverse assortment of contents. The model in [12] categorises customers based on their web browser, the number of pages they visit, the amount of time they spend on each page, or the search engine keywords they use; [14], [13] categorise customers based on their interest in specific product categories and the frequency with which they visit those categories.

When a consumer visits a website, [15] uses text mining techniques to determine the terms that are used the most frequently. These terms are then used to characterise the customer visit to the page. This method attempts to determine the user interests based on the content of the pages

that they view. It is also possible to construct a profile through the use of customer questionnaires [16] or a mix of purchasing data, demographic data, and personal information [17]. After the characterizations of the client have been determined, it is usual practise to utilise clustering methods in order to find groups of sessions that share a common behaviour or interest in the same topic. With the help of this data, we are able to modify both the content and the structure of the website. This is done for a variety of reasons, including, but not limited to, changing and personalising information; recommending products; analysing consumer behavior; and gaining insight into consumer preferences in specific products.

A second team makes use of other mining techniques in order to make predictions regarding user behaviour. Models that are based on the navigational sequences of the user [18] are able to anticipate the next action that the user will take. These models are shown through the use of Markov chains. There are a few drawbacks associated with these models, including the fact that it takes a significant amount of effort to construct them and that they can only react to short-term thinking. Nevertheless, each of these methods has its redeeming qualities. As shown in [19], these statistical models can be made more accurate by integrating several clustering methodologies with Markov chains. After the user sessions have been categorised through the use of a clustering method, a distinct Markov chain will be assigned to each individual cluster.

Techniques for process mining can be of assistance to you if you are looking for further information about what transpired throughout the experience that a user had. In the context of online retailing, the example provided by [20] demonstrates one such solution. For this reason, e-commerce websites should no longer make use of user behaviour models such as Petri nets and BPMNs. These models seek to depict user behaviour by employing a paradigm similar to that of a workflow. However, this approach is no longer valid. As a consequence of this, the methods that are utilised in [21] are restricted to occurrences that involve a high level of abstraction, which makes it challenging to recognise patterns that are associated with rare behaviours. Instead of presuming that the system adheres to a certain set of imperative criteria, a set of constraints is assumed, and anything that does not violate these constraints is allowed. The method in question is referred to as a declarative approach. The most important step here is to locate these boundaries [22] - [24]. These limits are frequently expressed through the use of temporal logic. The effort of describing properties can be made simpler for the analyst by using a collection of patterns that are based on common workflow structures and which are referred to as patterns in the Declare property description language [24]. Declare allows for the description of user sessions through the use of fundamental causality linkages. You also have the option of using MP-Declare [25], which extends the functionality of Declare patterns to include data and temporal constraints. In spite of the fact that it is restricted to a certain pattern set, it is not possible to examine universal formulas. In order for the analysts to test new patterns, they would need to establish their own instances of specific functions. Regardless of the patterns that are used, the temporal logic characteristics that have been stated need to have some kind of model checking tool used to validate them against the website log.

Logs from e-commerce websites can now be analysed with the help of powerful commercial solutions such as Google Analytics. Google Analytics is responsible for the management of network traffic, the collection of data regarding user sessions, and the generation of reports that summarise user behaviour. This traffic-based metric can also take into account the personal and geographical information of other users. However, Google Analytics does not support the importation of web server logs; nevertheless, page tagging techniques are utilised in order to capture data that may subsequently be analysed.

When compared to log-based analysis, these methods have a number of disadvantages, including the requirement that page tags be added to each and every page, the complexity of said page tags,

the possibility that users will notice a difference in the amount of time it takes to download your website, and concerns regarding the user privacy. On the other hand, Google reports contain a wealth of information that is exclusive to subject matter experts and can only be viewed by those individuals. makes a proposal for a strategy.

Proposed Method

This section provides an explanation of how the data from a blogging website was collected and analysed in order to discover patterns and produce insights into the website and the behaviour of its users.

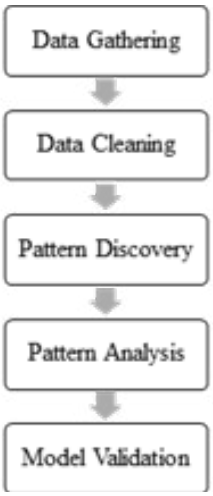

Fig. 2 Proposed Model

Data Gathering

In order to conduct web usage mining, a brand-new website has been created specifically for the purpose of analysing live online logs. The blog on this website contains posts, poems, videos, and other forms of media covering a wide range of topics. The website received visitors, and for a period of two weeks, logs were collected before being used for examination with a variety of search engine optimization tools and strategies.

Data Cleaning

The requests that were made to access the log files were saved using the combined log format, which was employed to do so. The requests that had the status code 200 and the method GET were the ones that were analysed and led to the discovery of patterns. In addition, the totality of all client requests that included images have been removed, including those whose filenames ended in .gif, .jpg, .ico, .png, and similar extensions. Because the objective is to analyse the behaviour patterns of visitors, requests made by robots and administrators have also been disregarded. Users are identified based on the IP addresses that are recorded in the access logs. Figure 3 depicts a fraction of the data that is clean and organised from the web server logs. This subset contains the noisy data that was discussed earlier.

Pattern Discovery and Analysis

After the data has been cleaned up, it will be much simpler to do an analysis of the user access patterns. Visualization tools allow even individuals without a technical background to profit from

the process of recognising patterns, trends, and popular pages. Instead of inspecting the access log files on the server, a data analysis and visualisation tool is used to search for patterns and trends. What kinds of questions can be addressed using data on access and referrals? How many different people have stopped by to look at this page in the past week? How many people hailing from a specific country access my website on a daily basis? The access logs and referral logs should be able to provide the answers to these inquiries. To begin with, both the structure and the content of the website need to be analysed in order to discover any patterns. According to a network of pages that are connected to one another in some way. On the other hand, outbound links are hyperlinks that take the user away from the current website in order to connect them with another. By making use of this information, a proprietor of a company can obtain a better understanding of the strengths and limitations of his website. This understanding is contingent on the frequency with which people view the website.

Model Validation

When traversing the pages of an e-commerce website, customers frequently use two distinct kinds of interactions: GET requests for data and POST requests for actions such as adding an item to their shopping cart, making a purchase, or checking in. A website log is where connected pieces of information, such as an IP address or the time that the contact took place, are logged. On every e-commerce website, there are recurring events that take place, such as customers browsing product categories and adding products to their shopping carts. As a result, it is feasible to give a universal strategy for classifying events found in web logs in accordance with the product classification. For the time being, we will discuss the proposed method for recognising significant events, recognising patterns of behaviour, and performing model checking based on the earlier classification to recognise significant occurrences.

When it comes time to test the model, we will be using temporal logic equations to represent events. This will enable us to view the log as a Kripke structure that represents the model that has to be analysed. During this time, which is referred to as the preprocessing phase, the log model will be developed.

Similar taxonomies have been proposed by different authors but including only main sections. From the homepage (level 0) different sections can be accessed (level 1). Two different types of sections can be distinguished.

Components of fundamental importance that are relevant to the products' primary classification. These are the sections where you can find every product. Although distinct product categories are the norm, this is not always the case. Sometimes they can overlap. On some e-commerce websites, the same item could be filed under several different categories at the same time.

The secondary parts of the website, which give a supplement to the primary classification of the website offers, allow users to access some of the products that are sold on the online store. As was the case with the scenario before this one, not all of the products are required to be made accessible through these supplementary categories. There are also two different kinds of supplemental sections, and the distinction between them is based on whether or not the components contained within them constitute permanent additions to the component. Examples of sections that contain transitory products include ones that have offers or sections with fresh products that are frequently replenished. Examples of supplementary sections that have permanent links are those that organise products according to manufacturer, subject, or other categories.

It is common practise to subdivide a section into many subcategories in order to more precisely categorise products. This is true regardless of the type of section being discussed. Each unique website that engages in e-commerce receives its own unique organisation (categories, levels, etc.).

Within the website navigational map, earlier sections are represented by specific web pages, and it is through these pages that users can go to products and other sections. Typically, a great number of additional links and menu options are provided in order to improve the user overall experience with the website. It is possible to bypass levels and integrate horizontal linkages across sections, making it feasible to access items directly from any level in the hierarchy, not only the one that is now active. For instance, e-commerce websites typically have search engines as a standard functionality, and users can employ these engines to look for certain goods. These mechanisms, which function in a manner analogous to that of supplementary sections in order to provide customers with an alternate means of accessing merchandise,

Let have a look at the web server logs to see if the most significant aspects of the website are reflected in them, now that we have defined those aspects.

Consider N∈N, which represents the entire number of levels in the system. The set of atomic propositions V = {vi|i1,2,…,N} (W = V = {wi|i1,2,…,N} can be used to categorise the many distinct types of events that can take place when a person enters the main (secondary) section of a building. There are a number of different types of events that can take place when a person enters the main (secondary) section of a building. For each of the events that correspond to a particular main (secondary) level-i section, the vi(wi) proposition will be noted.

The set of atomic propositions that are bijective with the main (secondary) sections of level i is denoted by the notation Mi(Si). This notation is used for each major section of each level. According to that, the event that corresponds to visiting the mj major part of level j≤N can be expressed as the logic formula vj - m1 -m2... mj, where m, M1 for each 11..j. In other words, the formula can be written as viewing the mj main section of level. In addition to being used for the primary components, the W and Si sets will also be utilised.

When it comes to products, the atomic proposition vp is used to explain the download process in its entirety (view product). An alternate hypothesis could be proposed in place of the current one in order to assist in the division of the products. On the other hand, this would make conducting an analysis challenging or even impossible, and the outcomes would be less interesting than when dealing with categories that are related to each segment.

By identifying several common events and distinct sets of common events, it is feasible to present pattern questions that are not dependent on the e-commerce website that is being investigated. As was said before, model checking can be utilised to research certain states, as well as their development and the connections between them. Using model checking, e-commerce web server logs can be analysed to determine which sections of a website are most popular with visitors, how users navigate to specific sites on the website, how the different sections relate to one another, and even which sections result in customers making actual purchases.

Secondary Categorization using Machine Learning

One example of a probabilistic classification technique is the Naïve Bayes (NB) classifier, which is derived from Bayes' theorem. When it comes to text classification, the performance of NB classifiers is superior to that of other machine learning algorithms. Nonetheless, NB classification does require the independence of the text features. The following is the formula for an NB classification classifier that is based on the Bayes theorem:

$$P(X \mid C_i) = \prod_{(k-1)}^{n} [\![p(x_k [\![| C]\!]_i)]\!] \quad (1)$$

The NB is superior to other machine learning algorithms in terms of the amount of time required to implement the algorithm as well as the amount of time required to train new data. As a direct consequence of this, the performance of this algorithm improves as the amount of training data increases.

Results and Discussions

In this section, over the period of a month, total request processed in the server on a website (Table 1) is found to be 10154 (5 MB). After the process of removal of irrelevant information, the total request available for the analysis is found to be 5154 (2.4 MB). For experimental purpose, a total of 10% of the log file contents are utilised for testing the machine learning to perform the pattern discovery analysis. The dataset is obtained from https://www.kaggle.com/competitions/h6751-text-and-web-mining/overview.

Table I. Website Statistics

Aspects	Information
Unique users	158
Total sessions	295
Average pages accessed/session	5
Average time spent/session	3m 53s

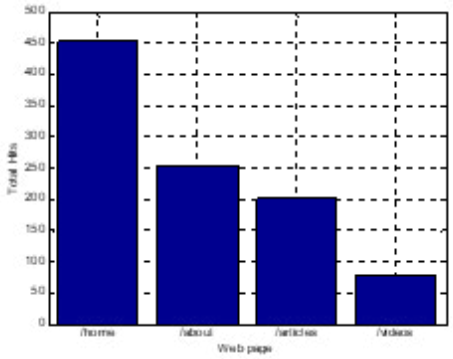

Fig. 3 Frequently Accessed Pages in a Website

Figure 3 presents an analysis of the pages that had the greatest number of views from site visitors. The analysis of these patterns can assist the proprietor of the website in determining which pages draw the attention of the visitors and which ones assist in identifying their preferences. When a person visits a website, it is possible to find out the previous page they were on.

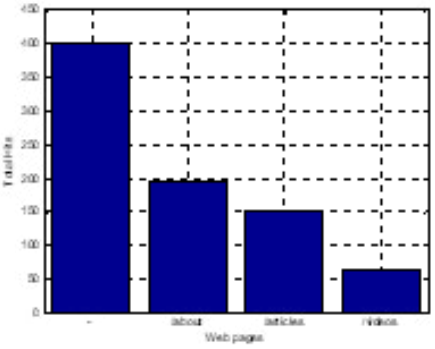

Fig. 4 Top Links of the website

As can be seen in Figure 4, consumers who used the symbol '-' imply that they typed the URL of the website straight into their web browser. By analysing these trends, it may be possible to determine which pages on the website maintain the interest of site visitors and, as a result, encourage them to explore the site further. Imagine websites that allow you to shop online as an analogue that is easier to understand. The site provides recommendations to site visitors that are determined by analysing the records of multiple customers. It is possible that the website will recommend that Customer A purchase additional accessories such as screen-guards and coverings, if previous customers have also purchased these kinds of items.

Because of the information that was acquired, it is now possible to make adjustments to the strategy and put new strategies into action in order to increase the exposure of the website. It would appear from this information that the Articles and Videos sections of the website were the ones that received the greatest number of visitors. These characteristics should be given priority on websites because they attract more visitors than other web page characteristics. In order to bring in a greater number of customers, it is recommended that advertising campaigns be developed based on the most visited pages of the website. It is vital to conduct an analysis of the behaviour of website users in order to maximise a website visibility and quickly meet the purpose and goals for which the website was created as seen in figure 5-7.

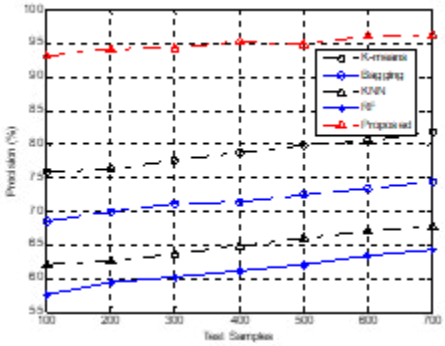

Fig. 5 Precision of Pattern Discovery Analysis with various Test Samples

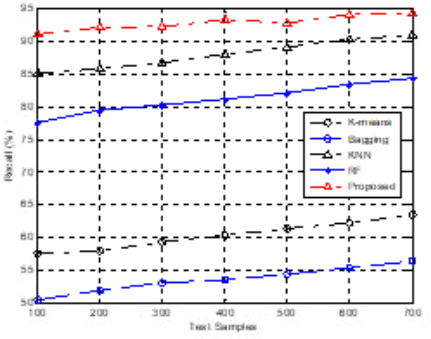

Fig. 6 Recall of Pattern Discovery Analysis with Various Test Samples

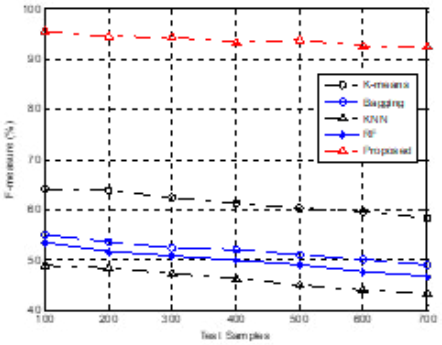

Fig. 7 F-measure of Pattern Discovery Analysis with Various Test Samples

Conclusions

This research aims to better understand how consumers interact with e-marketing websites in order to better target their purchases. The purpose of this research is to better understand how consumers interact with e-marketing websites. An algorithm for machine learning is utilised in order to perform the tasks of analysing log data from a large number of users during the training phase and giving user-specific relevant information during the testing phase. It is a test to determine whether or not machine learning can produce findings that are relevant to user behaviour, and the major criterion used to evaluate that capability is accuracy. According to the findings of this research, the machine learning model that is being considered for use is more accurate than other methods that are currently being used.

References

1. Bakariya, B. (2021). Efficient Approach of Analyzing and Generating Intrinsic Information from Weblog. National Academy Science Letters, 44(6), 525-527.
2. Chayanukro, S., Mahmuddin, M., & Husni, H. (2021, April). Understanding and assembling user behaviours using features of Moodle data for eLearning usage from performance of course-student weblog. In Journal of Physics: Conference Series (Vol. 1869, No. 1, p. 012087). IOP Publishing.
3. Siwach, M., & Mann, S. (2022). Anomaly detection for weblog data analysis using weighted PCA technique. Journal of Information and Optimization Sciences, 43(1), 131-141.
4. Om Prakash, P. G., Abdul Rahman, A., Nagaraj, J., & Sivakumar, N. (2022). Forecasting the User Prediction from Weblogs Using Improved IncSpan Algorithm. In Sustainable Communication Networks and Application (pp. 767-777). Springer, Singapore.
5. Agarwal, I. Y., Rana, D. P., Suri, K. R., Jain, P., Awasthi, S., & Roy, K. (2021). Behaviour Anomaly Detection With Similarity-Based Sampling for Imbalanced Data. In Data Preprocessing, Active Learning, and Cost Perceptive Approaches for Resolving Data Imbalance (pp. 177-194). IGI Global.
6. Vanitha, N., & Suriakala, M. (2021). A Thorough Study on Weblog Files and Its Analysis Tools. In Smart Computing Techniques and Applications (pp. 665-670). Springer, Singapore.
7. G Martín, A., Fernández-Isabel, A., Martín de Diego, I., & Beltrán, M. (2021). A survey for user behavior analysis based on machine learning techniques: current models and applications. Applied Intelligence, 51(8), 6029-6055.
8. Wang, C. (2021). Analysis of Students' Behavior in English Online Education Based on Data Mining. Mobile Information Systems, 2021.
9. Cui, Y., & He, Q. (2021). Inferring Twitters' Socio-demographics to Correct Sampling Bias of Social Media Data for Augmenting Travel Behavior Analysis. Journal of Big Data Analytics in Transportation, 3(2), 159-174.

10. Das, K., & Sinha, S. K. (2021). User Behaviour Analysis from Various Activities Recorded in Social Network Log Data. In Applications of Internet of Things (pp. 243-253). Springer, Singapore.
11. Mukunthan, B., & Arunkrishna, M. (2021, March). Spam Detection and Spammer Behaviour Analysis in Twitter Using Content Based Filtering Approach. In Journal of Physics: Conference Series (Vol. 1817, No. 1, p. 012014). IOP Publishing.
12. Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur. com. Expert Systems with Applications, 39(12), 11243-11249.
13. Cho, Y. H., & Kim, J. K. (2004). Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce. Expert systems with Applications, 26(2), 233-246.
14. Su, Q., & Chen, L. (2015). A method for discovering clusters of e-commerce interest patterns using click-stream data. electronic commerce research and applications, 14(1), 1-13.
15. Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Pérez, J. M., & Perona, I. (2013). Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. Expert Systems with applications, 40(18), 7478-7491.
16. Wu, R. S., & Chou, P. H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications, 10(3), 331-341.
17. Kim, K. J., & Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. Expert systems with applications, 34(2), 1200-1209.
18. Bernhard, S. D., Leung, C. K., Reimer, V. J., & Westlake, J. (2016, July). Clickstream prediction using sequential stream mining techniques with Markov chains. In Proceedings of the 20th International Database Engineering & Applications Symposium (pp. 24-33).
19. Lu, L., Dunham, M., & Meng, Y. (2005, August). Mining significant usage patterns from clickstream data. In International Workshop on Knowledge Discovery on the Web (pp. 1-17). Springer, Berlin, Heidelberg.
20. Poggi, N., Muthusamy, V., Carrera, D., & Khalaf, R. (2013). Business process mining from e-commerce web logs. In Business process management (pp. 65-80). Springer, Berlin, Heidelberg.
21. Maggi, F. M., Bose, R. P., & van der Aalst, W. M. (2012, June). Efficient discovery of understandable declarative process models from event logs. In International Conference on Advanced Information Systems Engineering (pp. 270-285). Springer, Berlin, Heidelberg.
22. Räim, M., Ciccio, C. D., Maggi, F. M., Mecella, M., & Mendling, J. (2014, October). Log-based understanding of business processes through temporal logic query checking. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 75-92). Springer, Berlin, Heidelberg.
23. Burattin, A., Cimitile, M., Maggi, F. M., & Sperduti, A. (2015). Online discovery of declarative process models from event streams. IEEE Transactions on services computing, 8(6), 833-846.
24. Burattin, A., Maggi, F. M., & Sperduti, A. (2016). Conformance checking based on multi-perspective declarative process models. Expert systems with applications, 65, 194-211.
25. Sturm, C., & Schönig, S. (2018, March). Big Data Meets Process Science: Distributed Mining of MP-Declare Process Models. In International Conference on Enterprise Information Systems (pp. 396-423). Springer, Cham.
26. Mehler, A., & Gleim, R. (2005). The net for the graphs–towards webgenre representation for corpus linguistic studies. WaCky, 191-224.