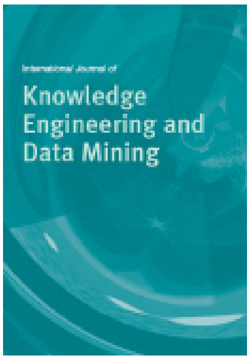
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Knowledge Engineering and Data Mining

ISSN online: 1755-2095 - ISSN print: 1755-2087
https://www.inderscience.com/ijkedm

Knowledge discovery in time series data with contextual event identification

Makarand Velankar, Vaibhav Khatavkar, Vinayak Jagtap, Parag Kulkarni

DOI: 10.1504/IJKEDM.2022.10049279

Article History:

Received:	09 February 2022
Last revised:	18 July 2022
Accepted:	27 June 2022
Published online:	10 October 2022

Copyright © Inderscience Enterprises Ltd.

Knowledge discovery in time series data with contextual event identification

Makarand Velankar*

MKSSS Cummins College of Engineering,
Pune, 411052 Maharashtra, India
Email: makarand.velankar@cumminscollege.in
*Corresponding author

Vaibhav Khatavkar and Vinayak Jagtap

College of Engineering,
Pune, 411030, Maharashtra, India
Email: vkk.comp@coep.ac.in
Email: vinayak5jagtap@gmail.com

Parag Kulkarni

iKnowlation Research Labs Pvt Ltd,
Pune, Maharashtra, India
Email: paragindia@gmail.com

Abstract: Massive temporal data generated in different domains need to be analysed for decision-making by various applications. This time-series data holds significant contextual knowledge in the form of hidden events. The need for automatic identification of such events is apparent. The lack of effective pattern identification techniques for contextual events suggests the need for efficient event identification methods for various applications. The study aims to propose a contextual event identification methodology in temporal data using exploratory data learning. The exploratory learning algorithm identifies appropriate uncertainty limits in an iterative approach to get desired information gain. Audio music streams and standard text datasets are used to test the method and retrieve contextual events. The result shows a 1.8% improvement for text data compared with an LDA approach and 0.04% improvement in the mean reciprocal rate for music data. Contextual event identification is helpful for different decision-making tasks in machine learning. The proposed system is extendable in different domains such as network, financial or medical, where event identification of temporal data is essential.

Keywords: data analysis; event discovery; exploratory learning; time-series data; uncertainty.

Reference to this paper should be made as follows: Velankar, M., Khatavkar, V., Jagtap, V. and Kulkarni, P. (2022) 'Knowledge discovery in time series data with contextual event identification', *Int. J. Knowledge Engineering and Data Mining*, Vol. 7, Nos. 3/4, pp.252–270.

Copyright © 2022 Inderscience Enterprises Ltd.

Biographical notes: Makarand Velankar received his PhD in Computer Engineering from Savitribai Phule Pune University. He has 11 years of industry experience and is associated with MKSSS'S Cummins College of Engineering for women, Pune, Maharashtra, India since 2001. His research interests include machine learning, AI, information retrieval, algorithms and computational musicology. He is a member of various professional bodies such as Computer Society of India, ACM, Indian Institute of Engineers. He is associated with International Society of Music Information Retrieval to promote research in computational music. He is also an entrepreneurship educator and approved mentor for start-ups.

Vaibhav Khatavkar completed his PhD from College of Engineering, Pune in 2022. He is actively involved in various social-technological transfer programs in India. He previously worked with Ascellent technologies from June 2007 to July 2008 as an exposure to the industry. He is presently working as an Assistant Professor at the Computer Engineering Department. His research interests include machine learning and natural language processing. He is working on industrial projects in the domain of natural language processing. During a springer conference in 2015, he won the best paper award for the paper titled 'Context vector machine for information retrieval' in 2015.

Vinayak Jagtap is a passionate researcher in the field of computer science. His research areas include data science, machine learning, and data security. He holds a Bachelor of Engineering from Walchand College of Engineering, Master's in Engineering from Pune Institute of Computer Technology, and currently pursuing his PhD from the College of Engineering Pune. He has a couple of patents filed along with dozen publications.

Parag Kulkarni is an entrepreneur, machine learning researcher and author of best-selling innovation strategy, ML and data science books. He holds a Bachelors from Walchand College of Engineering Sangli (1990), PhD from IIT Kharagpur (2001), management education from IIM Kolkata and was conferred higher Doctorate DSc by UGSM monarch, Switzerland (2010). He is the world's first higher doctorate in the area of knowledge innovation. He has published over 300+ research papers and articles in peer reviewed journals and renown conferences. He invented over a dozen patents and authored 14 books with world's best technical and business publishers like Bloomsbury, IEEE, Wiley, Prentice Hall, Springer, Oxford University Press, etc.

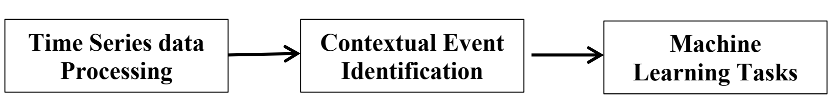
1 Introduction

Time series data are generated in various domains and is analysed to make appropriate decisions based on predictions. The prediction is based on the probabilistic approach using statistical models. Recent examples of COVID data analytics for an association of temperature and humidity prediction by Qi et al. (2020) or possible cases for medical facilities by Benvenuto et al. (2020) considering the possible spread are the examples of time series data where past daily data is used as a reference for analytics. Different parameters are considered in the statistical modelling, and the predictions are based on certain assumptions and available data. Inaccuracy in time series data collection or wrong assumptions can lead to failure of the proposed model, and wrong predictions are possible in such cases. For example, stock market analysis for predicting stock prices from Idrees et al. (2019) or time-series data for social media analysis performed by Ibrahim and Wang (2019) can be used to predict future decisions. Traffic data analysis in a computer network by Madan and Sarathi Mangipudi (2018) or power flow simulation as per Selim et al. (2018) helps predict the resource requirements at a specific time or day in the future. This time-series data is enormous in most applications and requires enormous computational resources and time.

Event-based analysis of time series data reduces the efforts in analysing the entire time-series data, and analysis of critical or extreme events is sufficient in many applications such as demand forecasting of vehicles for Uber performed by Laptev et al. (2017). This critical event discovery is referred to as contextual event identification (CEI) considering specific tasks. Contextual events in the same time series data may vary depending on the task. Each task will have different requirements, and the contextual event information required is likely to vary. Contextual events to predict the vehicle demand in specific time slots on working days will differ from contextual events to predict the demand in a specific scheme or festival. Some examples of CEI of specific tasks are bursty event identification for text streams proposed by Fung et al. (2005), Twitter event detection for topic summarisation by Cordeiro (2012), or epileptic event detection in EEG data done by Correa et al. (2019). Various models are proposed in the literature for analysis and prediction using time series data with autoregressive moving average or ARIMA is one of the most widely used approaches by researchers (Alsharif et al., 2019; Qin et al., 2017). Event discovery or identification is performed by exploring time series data and identification of specific patterns (Amornbunchornvej et al., 2018; Souto and Liebig, 2016; Zhou et al., 2017). Learning specific data patterns in temporal data is referred to as exploratory data learning (EDL). EDL approach is a data-driven approach, and data exploration is helpful for the inference in the context event identification. Some researchers have applied a data-driven approach for time series event identification in different domains, such as rainfall by Thiesen et al. (2019) or clinical data by Harutyunyan et al. (2019). Data-driven knowledge discovery is a promising approach, as Ding and Stirling (2017) mentioned. EDL investigates the temporal data to discover relationships between available contextual knowledge and unknown concepts. It discovers the events which are likely to be crucial from the task-related context and valuable for further processing. The identification of contextual events is a crucial and challenging task.

The primary motivation of the work is to propose a general methodology for CEI suitable for any time-series data, as shown in Figure 1. The event identification of time series data helps with standard tasks in machine learning such as classification as per Pandey et al. (2020), clustering as per Ives et al. (2016), and predictions. However, not all events or entire data are essential for the specific application, and events identified using contextual information are the main decision-making points in the vast time series data. Therefore, further processing of these decision-making points reduces the time and computing overheads and provides more accurate results.

Figure 1 Time series data processing



- contextual event identification using exploratory data learning (CEI-EDL) methodology for the time series data analysis proposed
- enhancement of machine learning tasks using contextual events with improved accuracies
- efficient learning with an EDL approach
- effective use of data uncertainty in CEI.

The paper is organised in the following manner. First, the literature survey covers knowledge discovery and event identification approaches used for time series data. Then, the proposed CEI-EDL method is explained with a mathematical foundation. Next, the experimental setup covers the experiments performed to test the methodology along with the results obtained. Finally, the conclusions and possible future scope are covered in the last section.

2 Literature survey

In general, numeric data in sequence of medical and other domains like moves in games or sports can be modelled as time series data. In addition, the structured and unstructured data can be modelled in a time series to discover hidden knowledge. Data mining and knowledge discovery are essential in most applications related to time series data. Researchers deploy different approaches to extract valuable knowledge from the vast time series data.

The knowledge discovery approaches used either application or data-centric. Self-organising feature maps to identify sleep-related disorders by Ultsch (1999) from multivariate time series data such as EEG, EMG, EOG, EKG, and symbolic knowledge represented using unification-based grammars algorithm was built on the Kullback-Leibler divergence using the probability distributions. The butions of the discretisation symbols were proposed by Morchen and Ultsch (2005) for the optimised discretisation to transform numeric time series into knowledge in the symbolic form. Computational theory of perception was used for the information-theoretic fuzzy approach by Last et al. (2001) knowledge discovery for time-series databases of stock market, and weather. For the sentiment knowledge discovery in Twitter streaming data, Bifet and Frank (2010) proposed a sliding window Kappa statistic computation in time-changing data streams.

During a detailed survey performed by Ramirez-Gallego et al. (2017) for data stream mining, they expressed a need to develop new and more sophisticated data preprocessing methods, one of the critical stages of the knowledge discovery from complex data types. Some researchers used an integrated approach to discover knowledge. For knowledge discovery, collaborative and distributed approaches to computing were used (Singh et al., 2007; Pekkola et al., 2013). Event identification is a critical knowledge discovery step in most applications based on time series as it leads to a basis for critical decision making.

The most challenging part of sizable time-series data is identifying interesting events (Lan and Ma, 2008) or regions of interest. The LeadLine is one of the event identification interactive visual analytical situations, which integrates topic modelling, event detection (identification), and name data entity relationship based on 4W’s (Who, What, When, Where). The knowledge is mined using time sequence referencing and visually represented to extract more knowledge for decision making. They were represented as entity graphs and Geo-mapping based on the position of the context in the text (Cerbin and Kopp, 2006). Event detection was considered the crucial pre-processing step as knowledge discovery as per Guralnik and Srivastava (1999). A mathematical representation of event detection for time-series inputs was given by Kom et al. (1997). Temporal abstraction was proposed for understanding and interpreting time-series data in earlier research by Shahar et al. (1996). A unifying framework for detecting outliers and change points was proposed by Yamanishi and Takeuchi (2002) for non-stationary time series data. During the research in the temporal pattern mining framework, identifying foretelling patterns for observing and event discovery problems in multifaceted multivariate time series data, experimentation was performed by Batal et al. (2012) on the healthcare data of 13,558 diabetic patients, and functional patterns were identified for spotting and analysing adverse medical situations with diabetic patients.

Various approaches to specific application domains were adopted to find exciting events (Hunter and McIntosh, 1999; Akoglu and Faloutsos, 2010; Wu et al., 2018). An event in time series data has different interpretations across different domains but has a common goal of knowledge discovery. For example, neural networks were used successfully in nuclear power plants to find abnormal events (Ohga and Seki, 1993) and forecast coolant accidents (Radaideh et al., 2020). In the medical field, for event discovery of time-series data, EEG signals were processed to find events, for knowledge discovery (Tsien, 2000; Lawhern et al., 2013). In the domain of social networks, knowledge based on the post-event was determined using text or other multimedia inputs (Atefeh and Khreich, 2015; Cordeiro and Gama, 2016). For the finance sector, events were proposed based on transactions, notifications, and news as inputs by Escher (2008). Event-based approaches were also explored for specific time series data sets such as UCR and MPEG-7 Core Experiment CE-Shape-1 (Koknar Tezel and Latecki, 2011), and multivariate medical time series (Dua et al., 2011). Architecture for emergency event prediction using LSTM recurrent neural networks was proposed by Cortez et al. (2018). Cekinel and Karagoz (2022) the proposed event prediction from news text using subgraph embedding and graph sequence mining. Real-time event detection and classification in social text streams using an embedding approach was proposed by Singh et al. (2022). A detailed survey for disaster event detection from text data by Gupta et al. (2022) provides different techniques used for event detection. Modelling of sequential listening behaviours with attentive temporal point processing for event detection was proposed by Wang et al. (2021) for music recommendation. Various approaches were discussed to identify events from different domains for knowledge discovery which will help decision-making.

The exhaustive literature survey indicates that most existing systems use the entire time series for the processing which is a complex and tedious task for knowledge discovery. The existing systems mainly utilise the dynamic changes in the time series data as a basis for event identification. These systems are likely to ignore important events and may fail to discover hidden knowledge. Furthermore, these approaches typically miss complete information (hidden knowledge) in the time series data. Finding out uncertainties between input sequences helps to reduce such a loss of information. Hence, uncertainty-based event identification and successive machine learning approaches will undoubtedly improve performance. Hence, the hypothesis is proposed as "The generic methods used in time series event identification fail to capture hidden knowledge, which leads to incorrect decision making. CEI with different uncertainty bounds in the proposed system represents hidden contextual information. If uncertainty-based event identification with EDL is performed on time series to extract relevant and useful information, then this information helps in an accurate classification and clustering". As a hypothesis, a CEI-EDL method is proposed, and experiments are performed as illustrated in this paper.

3 CEI-EDL method

The proposed CEI-EDL method is inspired by the information gain theory proposed by Shannon (1948). As per Shannon, entropy is a measure of uncertainty for some variables or events. The more significant is the uncertainty, the greater the Shannon entropy. Shannon proposed that information reduces uncertainty and therefore reduces entropy. The proposed method uses uncertainty based on the Shannon theory for CEI of the primary events. EDL based on a data-driven approach does not rely on domain-specific information and is a more promising approach developed recently, suitable for various machine learning applications.

3.1 Mathematical foundation

The mathematical foundation used for CEI is based on information theory. Various symbols were used and their description is given below.

T_s Time series data with *di* as data value at time instance *ti* with *n* samples

S_j Segment j in *Ts* with total *m* segments

E_k Event *k* identified in segment *j*

U_e Total number of unique events in segment *j*

$F(E_k)$ Frequency of occurrence of event *Ek*.

Entropy *En* and information gain of specific event *Ig*(*Ek*) is calculated using the following equations (1) and (2), respectively.

$$En = (1/Ue) * \log * (Ue) \tag{1}$$

$$Ig(Ek) = F(Ek) * En \tag{2}$$

Uncertainty of specific event *Ek* is calculated as *Un*(*Ek*) using equation (3).

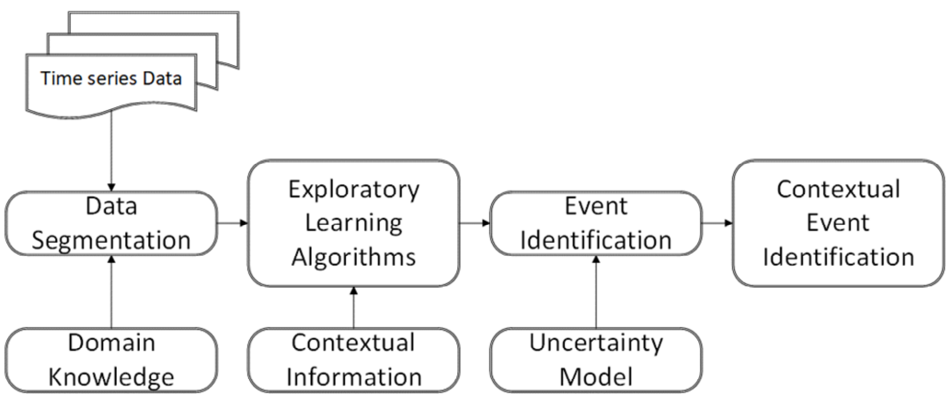
$$Un(Ek) = 1 - Ig(Ek) \tag{3}$$

Various researchers use this mathematical foundation proposed by Shannon for different applications, and this proven mathematical concept is applied for CEI of time series data.

3.2 Contextual event identification

Contextual events are identified using the uncertainty desired for the specific application. The threshold value or range of uncertainty is used for CEI. The basic functional blocks of the proposed method are shown in Figure 2. The segmentation of the time series data provides logical segments based on the domain information available. When domain information is unavailable or logical segmentation is not desired, the entire sequence is considered a single segment. Exploratory learning algorithms identify important events based on the contextual information desired. Primary events representation of the raw time-series data represents the semantic information associated with it. Uncertainty (Un) associated with each event (Ek) is computed using the basic mathematical formula for Un (Ek). Contextual events are identified based on the range/ threshold desired for Un (Ek) for the specific task. The range or a threshold is initially set with a heuristic approach and is modified as the model matures with more training data. The contextual events selected from the set of primary events are crucial, and it reduces further processing time with a focus on only contextual event information. Typical tasks such as time series comparison, classification, clustering, and predictions can be achieved using identified contextual events. These machine learning tasks using the CEI-EDL method is time-efficient and more accurate due to the semantic association than raw time series data.

Figure 2 CEI-EDL method



Event identification of time-series data generally needs domain knowledge and inputs from domain experts at the initial stage, and it helps in the possible segmentation and validation of contextual events for the task. For example, given an ECG time series data, what to look for at the critical events in the desired task and its further interpretation requires medical knowledge. The critical event can be a combination of primary events or primary events themselves, and the n-gram model is helpful for the grouping of the primary events in an ordered manner. The uncertainty is identified for these n-grams using the formulas mentioned in the previous subsection. The contextual events are identified using the uncertainty range or the information gain expected for the specific application. This EDL algorithm is data-driven and uses the contextual info to identify critical events that are contextually important for the task.

3.3 Exploratory learning

The proposed exploratory learning algorithm is data-driven and matures with the testing samples provided. The primary outcome of an exploratory learning algorithm is uncertainty limits for specific applications for CEI. The variety of test cases is essential in training the machine learning algorithm. The algorithm's accuracy depends on the quality of training data and domain expert knowledge involved in the initial stage. The algorithm adjusts the limits to provide the desired output for test cases used for training, and it is further tested for unknown samples to major the effectiveness in identifying the context-specific events.

Algorithm 1 Exploratory learning for uncertainty limits identification

Exploratory learning algorithm

  Input:

    Time series data samples with contextual events

    Uncertainty limits initialisation

  Output:

    Proposed uncertainty limits for a specific task

Procedure begin

  While the end of the training sample performs steps 1 to 6:

  1 *Read training time series data sample*

  2 *Perform segmentation of time series data sample (s: number of segments)*

  While the end of s performs steps 3 to 5:

  3 *Identify the contextual event with uncertainty limits*

  4 *Verify the identified contextual events*

  5 *If desired, adjust uncertainty limits to capture contextual events in the segment*

  Identify optimum uncertainty limits for training sample n

  6 *Store limits for sample in the table*

  While the end of limits in the table performs steps 7 to 11:

  7 *Read limit*

  8 *Read training time series data sample*

  9 *Perform segmentation of time series data sample (s: number of segments)*

  10 *Identify the contextual event with uncertainty limits*

  11 *Find the identified contextual events for a specific limit*

  Find optimum uncertainty limits for all training samples

  12 *Test the uncertainty limits for testing data*

  13 *If the desired outcome is satisfactory, then stop*

    Else increase training and test samples and continue

Procedure end

After several iterations of the algorithm, it is most likely that appropriate uncertainty limits will be obtained. If the problem persists, one needs to modify parameters such as more samples, the segmentation method used, and the accuracy expected to get desired results from the data. Since this approach is data-driven, the more significant number of samples and coverage of maximum variety in data samples will likely provide better results. The decision regarding the number of samples to be used, segmentation, and accuracy expected are highly dependent on the desired task. For critical applications like healthcare, it is desired to have more samples with higher accuracy expectations as the decision based on the outcome may be crucial. For a non-critical application like text or audio streaming, it may not be that crucial, and if the algorithm can provide expected output and reduces human efforts with time efficiency for CEI, it is acceptable for specific tasks.

3.4 Experimental methods

Experimentation is carried out for news feed as text data and music streaming as audio data to identify the contextual events using the proposed CEI-EDL method. For textual data, logical segments are sentences, and primary events can be words or sets of words in text data. In audio musical data, silence is used for the melodic segmentation, and the musical sequences of notes can be melodic events. The sample small duration time series data are represented to explain the implementation of the proposed method for simple tasks. Primary events and contextual information are associated with the task and the domain. For news feed, meaningful words or sets of words provide context and contextual events in words or sets of words based on the desired uncertainty. For music streaming, the musical notes or some sequence of notes provides meaningful musical context, and the contextually applicable events are discovered from the primary events considering computational music-related tasks.

4 Newsfeed text data

The majority of work related to text data mining uses machine learning and pattern recognition-based approaches for knowledge discovery. In systematic reviews by Usai et al. (2018) and Sirichanya and Kraisak (2021), the authors have presented various approaches used for data mining. An evolutive frequent pattern tree-based incremental knowledge discovery algorithm proposed by Liu (2022) provided improved results for online comments and news data. Associative rule-based approach of Al-Radaideh and Al-Khateeb (2015), context-based approach by Sonawane and Kulkarni (2016), multi-perspective machine learning approach by Khatavkar et al. (2019), agent mining with ontology by Grislin-Le Strugeon et al. (2021) integrated intelligent approaches involving machine learning by Tian et al. (2021) are typical approaches used for text data mining. The importance of text pre-processing in text mining and its impact on results is explored by Hickman et al. (2022). Modelling text data as time series is not a novel concept and it was explored by researchers in event identification based on Landau et al. (2009) and the LDA model on topic selection by Lee and Kim (2018). Recently a novel application to identify sarcasm in textual data is explored by Mehndiratta and Soni (2019). The facts in text data carry more significance when temporal information is added in it. For example, the terms 'President', 'USA' and 'Bill Clinton' will be more significant in an event when year scope '1993–2001' is added to it by Ling and Weld (2010). Li et al. (2022) in their work has applied temporal data from text for climate prediction using deep neural network.

For the newsfeed data, the time at which the news is generated and read will provide temporal information. The sequence in which the news is read will provide the context for the terms in the news article. The certainty of an event thus will change on revealing the context as time proceeds. Thus the aim of the experiments presented below is to calculate the certainty of the events in the news at a particular instance of time.

Text data are collected from the standard BBC news dataset (2006). The dataset consists of 2,225 news articles with five various categories, namely, business, entertainment, politics, sport or tech. An example article related to sports cricket is presented to explain the CEI-EDL method for identifying contextual events from the news article based on EDL and uncertainty of data. Table 1 presents the unique events identified with unigrams (UE), word count (WC), uncertainty (Un), and information gain (Ig) computed per event using the formulas mentioned earlier. Table 1 values are computed using the Entropy as 0.04106, computed from equation (2).

Table 1 Events and Information Gain Computation for unigrams

EventName	*UE*	*WC*	*Un*	*Ig*
flintoff	1	20	0.1786	0.8213
play	1	12	0.5071	0.4928
fit	1	9	0.6303	0.3696
#	9	6	0.7535	0.2464
#	13	4	0.8357	0.1642
#	1	3	0.8767	0.1232
#	14	2	0.9178	0.0821
#	141	1	0.9589	0.0410

Note: # list of events

Table 1 show that the event 'Flintoff' can identify the document with the highest information gain as 0.8213 and is a context-specific event in this document considering unigrams. The uncertainty of the events increases as the frequency of events in the document decreases. For events with frequency 1, the uncertainty is more. Thus, 141 events are not providing any relevant information about this document. This approach is similar to the term frequency-inverse document frequency (TF-IDF) approach used by many researchers such as Li and Lin (2017) or Guo et al. (2019). The main difference here is the use of n-grams for the context of specific event identification, which relies on information gain associated with a combination of unigrams.

Unique events identified with bigrams (entropy as 0.0351) and trigrams (entropy as 0.03485) along with Un and Ig calculations are presented in Tables 2 and 3, respectively.

Table 2 Events and information gain computation for bigrams

Un	*Ig*	*bigrams*	*bigram frequency*
0.8595	0.1404	5	4
0.9297	0.0702	3	2
0.9648	0.0351	213	1

The bigrams and trigrams provide the temporal info significant in the text data. The 5 bigram events {michael, vaughan}, {fit, play}, {cape, town}, {defeat, cape}, {marcus, trescothick} having uncertainty 0.8595 represents the first row of Table 2. After considering the uncertainty associated, it can be concluded that these events have some association with the context of the document. Further, the computation for trigrams revealed results, as shown in Table 3.

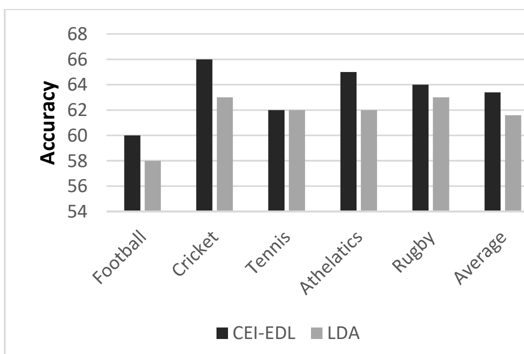
Table 3 Events and Information Gain Computation for trigrams

Un	*Ig*	*trigrams*	*trigram frequency*
0.9302	0.0697	1	2
0.9651	0.0348	223	1

The trigram event {defeat, cape, town} denotes the first row in Table 3, and the association with context is further reduced. Considering the uncertainty obtained in trigrams, it was not necessary to compute any n-grams beyond trigrams. A similar observation was observed in almost all documents of the dataset. Thus, we can conclude that possible essential events are likely to be identified using unigrams, bigrams, and max trigrams. Further, the association of the contextual events identified using the CEI-EDL method is computed and used to identify contextual events associated with the document.

The contextual events identified were further tested with the existing human-annotated titles of the news articles, and it was observed that 63.4 % of words matched the existing titles. Thus, it can be concluded that the contextual events identified provide 63.4% accuracy for the BBC sports news standard dataset with 786 documents spread across five categories, namely football, cricket, tennis, athletics, and rugby.

Figure 3 Comparison of CEI-EDL with LDA for BBC sports dataset



Theme or topic identification of the documents is a challenging AI problem. Latent Dirichlet allocation (LDA) is considered one of the successful approaches for topic identification in the document. It is an unsupervised machine learning approach and discovers topics using the linguistic model. The results of contextual events identified using the CEI-EDL method proposed are compared with the LDA approach for the standard BBC news dataset (2006), and the results are shown in Figure 3. The result shows an average improvement of 1.8 % over average LDA accuracy (61.6%). Further, it was observed that 24% bigrams and 7% trigrams obtained from documents contributed to CEI using the CEI-EDL approach and contributed to improved accuracy.

5 Music stream audio data

Time series modelling of music data is being used by researchers for many applications such as semantic music annotation by Coviello et al. (2010), motif identification by Rao et al. (2014), classification by Sharma et al. (2021), music storage and retrieval by Si (2021), audio event classification by Hershey et al. (2021), video to music recommendation by Pretet et al. (2022). The music stream is modeled as time series data for outlier detection by Herskind Sejr et al. (2021) and proposed a need for closer interaction with domain experts and users. Music audio stream with one channel is single-dimensional time series data, whereas the sound recorded for two or more channels is multidimensional time-series data representing each channel recording. Multichannel recording can be converted to a mono channel with a one-dimensional data stream. Musical facets such as pitch, intensity, timbre, and rhythm are associated with the interpretation of music by humans. The mono channel music stream is used as an input for the experiments performed for the melodic event identification. Pitch information is the crucial component of the melody of the music. Exploratory learning is performed for the time-series data to identify associated pitch information present in the data stream. The pitch information represents the original data in the semantic information related to acoustics. These pitch values are associated with musical notes. This sequence of notes representing melody is a melodic representation of the music audio stream. The notes or specific sequence of notes is primary events for the melodic representation.

Contextual events are identified in the query by the humming (QBH) application. In QBH, the user attempts to hum the song melody possible, such as yodelling or using consonant syllables like da-da or la-la to retrieve the required song from the music database. This content-based retrieval application needs to convert the humming melody to a possible note sequence to match the relevant match in the database. The note sequence representation required is at the broader level by ignoring the music ornamentation part, and the database of music is represented using melody at the broader level.

The following example illustrates the proposed CEI-EDL method for discovering the contextual events, which are a sequence of prominent notes hummed for pattern matching in the database. The sample dataset of a few amateur and seasoned singers was used to hum specific songs and generate the melodic sequences. The experimentation was performed for 30 samples, and the resultant melodic sequence is compared with varying uncertainty to identify the range of uncertainty desired. Since a sequence of notes is essential, bigrams (2-note sequence) and trigrams (3 note sequence) are used as possible contextual events. Further, these prominent note patterns are compared with the song patterns stored in the database to identify the distance and predict the possible songs with ranking. The song patterns stored in the database for 500 songs are used for comparison (Music Dataset MER 500, 2020). The dataset is available on kaggle for the usage for researchers. The data set consists of songs from five emotional classes, namely devotional, sad, happy, romantic and party. The following example illustrates the processing of primary events to identify contextual events for the QBH task in music information retrieval.

Primary events captured by the system in the form of the note sequence in the sample humming of one line of a song are represented in the form of note sequence as:

> C3 C3 C3 C3 B3 B3 B3 B4 C5 C5 C5 C5 C5 C5 C#5 C5 C5 C5 C5 C5 C5 A#4 A#4 G#4 F4 F4 D4

Tables 4, 5, and 6 represent the unigram, bigram, and trigram melody events provide the temporal info significant as total occurrences (O), a total number of events (E) identified along with their uncertainty computed respectively for the above sequence.

Table 4 Unigram melody events

O	*E*	*Ig*	*Un*	*Unigram events*
12	1	4.2265	–3.2265	C5
4	1	1.4088	–0.4088	C3
3	1	1.0566	–0.0566	B3
2	2	0.7044	0.2955	A#4, F4
1	4	0.3522	0.6477	B4,C#5,G#4, D4

Table 5 Bigram melody events

O	*E*	*Ig*	*Un*	*Bigram events*
10	1	2.7195	–1.7195	C5 C5
3	1	0.8158	0.1841	C3 C3
2	3	0.5439	0.4560	B3 B3, A#4 A#4, F4 F4
1	9	0.2719	0.7280	

Table 6 Trigram melody events

O	*E*	*Ig*	*Un*	*Trigram events*
8	1	1.9235	–0.9235	C5 C5 C5
2	1	0.4808	0.5191	C3 C3 C3
1	15	0.2404	0.7595	

The contextual events in the respective event column are marked in bold. These events represent the prominent sequences of events and help match the patterns in the databases. The events are further transformed into relative patterns based on reference notes for effective pattern matching. It can be easily observed that uncertainty increases with an increase in the n-grams, and the events with less uncertainty are contextual events to the QBH application. The overall cutoff for the average uncertainty makes unigram the first choice to identify the contextual event. The prominent temporal sequence representing the entire sequence in the compressed form is represented using unigram contextual events is C3 B3 C5 A#4 F4 in the example referred to. The contextual events with uncertainty, less than specific cutoff in bigrams and trigrams, are used to identify partial matching and ranking the songs in the order considering the matching. The choice of cutoff for the uncertainty is initially set with sample experimentation, and the data-driven exploratory learning approach helps it mature with a more significant number of queries and results obtained.

The results are tested using the MRR (mean reciprocal rank) measure, which is used to determine the accuracy of the retrieval task using ranking (Velankar and Kulkarni, 2018). The MRR is calculated in the following manner. For the specific humming queries, the intended song ranking was obtained after pattern matching comparison using contextual events. As a sample case for three queries, if the intended song appears at positions 1, 2, and 4, respectively, MRR is (1/1 +1/2 + 1/4) / 3, which is 0.5833. MRR varies in the range of 0 to 1, and the maximum the MRR, the better the retrieval. MRR values for 30 queries for 500 song database are computed using primary events and Contextual events identified. The MRR of 0.63 obtained using the CEI-EDL method provides better results with more accuracy than primary events with the MRR value of

0.59. The proposed method provides better results and is time-efficient as the number of patterns used for comparison, in the extensive databases of songs influences the time required. The CEI provides events with the most prominent patterns and is significantly less than primary events. The use of the n-grams with cutoff uncertainty values further improves the accuracy of the system.

6 Conclusions and future scope

The contextual event information obtained using the proposed CEI-EDL method represents the hidden knowledge, which is helpful in decision making. An EDL algorithm limits the uncertainty, which makes more precise and well-informed decisions. Depending on the application and decision support needed, one can use relevant contextual events. For example, N-gram selection is optimally made with the help of information gain. Furthermore, the contextual events identified in the time series data will be helpful for machine learning tasks such as classification, clustering, prediction, and recommendation.

For the standard BBC news dataset (2006), the CEI-EDL method has the edge over the conventional LDA approach for contextual event discovery. The exploratory learning approach for identifying bigrams and trigrams based on the uncertainty model contributed to the improvement. The method tested for music streams for QBH application provided contextual events suitable for the full or partial match for content-based music retrieval. It is time-efficient compared to the use of primary events for the task with more accurate results using MRR as a comparing parameter for information retrieval. One can further improve outcomes by training the system with an existing human-annotated context. More training data and the incorporation of an incremental learning approach will make the algorithm more robust. The context-driven system proposed is extendable for various applications related to gaming, stock market analysis, healthcare data, network analysis, and other time series applications. It will enhance the existing expert systems with hidden knowledge discovery. Uncertainty integration in a restricted form could be a future plugin to various expert systems for enhanced decision-making capabilities.

References

Akoglu, L. and Faloutsos, C. (2010) 'Event detection in time series of mobile communication graphs', in *Army Science Conference*, December, Vol. 1 [online] https://www.andrew.cmu.edu /user/lakoglu/pubs/EVENTDETECTION_AkogluFaloutsos.pdf (accessed 20 July 2020).

Al-Radaideh, Q.A. and Al-Khateeb, S.S. (2015) 'An associative rule-based classifier for Arabic medical text', *International Journal of Knowledge Engineering and Data Mining*, Vol. 3, Nos. 3–4, pp.255–273.

Alsharif, M.H., Younes, M.K. and Kim, J. (2019) 'Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea', *Symmetry*, Vol. 11, No. 2, p.240.

Amornbunchornvej, C., Brugere, I., Strandburg-Peshkin, A., Farine, D.R., Crofoot, M.C. and Berger-Wolf, T.Y. (2018) 'Coordination event detection and initiator identification in time series data', *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 12, No. 5, pp.1–33.

Atefeh, F. and Khreich, W. (2015) 'A survey of techniques for event detection in twitter', *Computational Intelligence*, February, Vol. 31, No. 1, pp.132–64, https://doi.org/10.1111/coin.12017.

Batal, I., Fradkin, D., Harrison, J., Moerchen, F. and Hauskrecht, M. (2012) 'Mining recent temporal patterns for event detection in multivariate time series data', in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 12 August, pp.280–288, https://doi.org/10.1145/2339530.2339578.

BBC News Dataset (2006) [online] http://mlg.ucd.ie/datasets/bbc.html (accessed 15 June 2022).

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S. and Ciccozzi, M. (2020) 'Application of the ARIMA model on the COVID-2019 epidemic dataset', *Data in Brief*, April, Vol. 29, p.105340.

Bifet, A. and Frank, E. (2010) 'Sentiment knowledge discovery in twitter streaming data', in *International Conference on Discovery Science*, 6 October, pp.1–15, Springer, Berlin, Heidelberg [online] https://link.springer.com/chapter/10.1007/978-3-642-16184-1_1 (accessed 15 June 2022).

Cekinel, R.F. and Karagoz, P. (2022) 'Event prediction from news text using subgraph embedding and graph sequence mining', *World Wide Web*, pp.1–26 (accessed 15 June 2022).

Cerbin, W. and Kopp, B. (2006) 'Lesson study as a model for building pedagogical knowledge and improving teaching', *International Journal of Teaching and Learning in Higher Education*, Vol. 18, No. 3, pp.250–7 [online] https://eric.ed.gov/?id=EJ1068058 (accessed 15 June 2022).

Cordeiro, M. (2012) 'Twitter event detection: combining wavelet analysis and topic inference summarization', in *Doctoral Symposium on Informatics Engineering*, January, Vol. 1, pp.11–16.

Cordeiro, M. and Gama, J. (2016) 'Online social networks event detection: a survey', in *Solving Large Scale Learning Tasks, Challenges and Algorithms*, pp.1–41, Springer, Cham [online] https://link.springer.com/chapter/10.1007/978-3-319-41706-6_1 (accessed 15 June 2022).

Correa, A.G., Orosco, L.L., Diez, P. and Leber, E.L. (2019) 'Adaptive filtering for epileptic event detection in the EEG', *Journal of Medical and Biological Engineering*, Vol. 39, No. 6, pp.912–918.

Cortez, B., Carrera, B., Kim, Y.J. and Jung, J.Y. (2018) 'An architecture for emergency event prediction using LSTM recurrent neural networks', *Expert Systems with Applications*, Vol. 97, pp.315–324, https://doi.org/10.1016/j.eswa.2017.12.037.

Coviello, E., Chan, A.B. and Lanckriet, G. (2010) 'Time series models for semantic music annotation', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 5, pp.1343–1359.

Ding, Y. and Stirling, K. (2017) 'Data-driven discovery: a new era of exploiting the literature and data', *Journal of Data and Information Science*, Vol. 1, No. 4, pp.1–9.

Dua, S., Saini, S. and Singh, H. (2011) 'Temporal pattern mining for multivariate time series classification', *Journal of Medical Imaging and Health Informatics*, 1 June, Vol. 1, No. 2, pp.164–9 [online] https://pubmed.ncbi.nlm.nih.gov/23921002/ (accessed 15 June 2022).

Escher, R.E. (2008) *Method of Providing a Financial Event Identification Service*, 23 December, Inventor; Recognia Inc, Assignee, United States patent US 7,469,226.

Fung, G.P.C., Yu, J.X., Yu, P.S. and Lu, H. (2005) 'Parameter free bursty events detection in text streams', in *Proceedings of the 31st International Conference on Very Large Data Bases*, pp.181–192.

Grislin-Le Strugeon, E., de Oliveira, K.M., Zekri, D. and Thilliez, M. (2021) 'Agent mining approaches: an ontological view', *The Knowledge Engineering Review*, p.36.

Guo, J., Mu, Y., Xiong, M., Liu, Y. and Gu, J. (2019) 'Activity feature solving based on TF-IDF for activity recognition in smart homes', *Complexity*, https://doi.org/10.1155/2019/5245373.

Gupta, A., Rani, M. and Kaushal, S. (2022) 'Disaster event detection from text: a survey', in *Computational Intelligence in Data Mining*, pp.281–293, Springer, Singapore.

Guralnik, V. and Srivastava, J. (1999) 'Event detection from time series data', in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.33–42, https://doi.org/10.1145/312129.312190.

Harutyunyan, H., Khachatrian, H., Kale, D.C., Steeg, G.V. and Galstyan, A. (2019) 'Multitask learning and benchmarking with clinical time series data', *Scientific Data*, Vol. 6, No. 1, pp.1–18.

Hershey, S., Ellis, D.P., Fonseca, E., Jansen, A., Liu, C., Moore, R.C. and Plakal, M. (2021) 'The benefit of temporally-strong labels in audio event classification', in *ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June, IEEE, pp.366–370.

Herskind Sejr, J., Christiansen, T., Dvinge, N., Hougesen, D., Schneider-Kamp, P. and Zimek, A. (2021) 'Outlier detection with explanations on music streaming data: a case study with Danmark Music Group Ltd.', *Applied Sciences*, Vol. 11, No. 5, p.2270.

Hickman, L., Thapa, S., Tay, L., Cao, M. and Srinivasan, P. (2022) 'Text preprocessing for text mining in organizational research: Review and recommendations', *Organizational Research Methods*, Vol. 25, No. 1, pp.114–146.

Hunter, J. and McIntosh, N. (1999) 'Knowledge-based event detection in complex time series data', in *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision-Making*, 20 June, pp.271–280, Springer, Berlin, Heidelberg [online] https://link.springer.com /chapter/10.1007/3-540-48720-4_30.

Ibrahim, N.F. and Wang, X. (2019) 'Decoding the sentiment dynamics of online retailing customers: time series analysis of social media', *Computers in Human Behavior*, July, Vol. 96, pp.32–45, https://doi.org/10.1016/j.chb.2019.02.004.

Idrees, S.M., Alam, M.A. and Agarwal, P. (2019) 'A prediction approach for stock market volatility based on time series data', *IEEE Access*, January, Vol. 7, pp.17287–17298, DOI: 10.1109/ACCESS.2019.2895252.

Ives, Z.G. (2016) 'Technical perspective: k-shape: efficient and accurate clustering of time series', *ACM SIGMOD Record*, Vol. 45, No. 1, pp.68–68.

Khatavkar, V., Velankar, M. and Kulkarni, P. (2019) 'Multi-perspective analysis of news articles using machine learning algorithms', *International Journal of Computer Applications*, February, Vol. 975, p.8887, http://dx.doi.org/10.5120/ijca2019918503.

Koknar Tezel, S. and Latecki, L.J. (2011) 'Improving SVM classification on imbalanced time series data sets with ghost points', *Knowledge and Information Systems*, 1 July, Vol. 28, No. 1, pp.1–23 [online] https://link.springer.com/article/10.1007/s10115-010-0310-3.

Kom, E.L., Graubard, B.I. and Midthune, D. (1997) 'Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale', *American Journal of Epidemiology*, 1 January, Vol. 145, No. 1, pp.72–80, https://doi.org/10.1093/oxfordjournals.aje.a009034.

Lan, Q. and Ma, C.A. (2008) 'Method of discovering patterns to predict specified events from financial time series', in *Proceedings of the 4th International Conference on Natural Computation*, pp.18–22, https://doi.org/10.1109/ICNC.2008.246.

Landau, D., Feldman, R., Lipshatz, O., Aumann, Y., Shemtov, H. (2009) *Method and System for Displaying Time-Series Data and Correlated Events Derived from Text Mining*, Reuters Ltd, assignee, United States patent US 7,570,262.

Laptev, N., Yosinski, J., Li, L.E. and Smyl, S. (2017) 'Time-series extreme event forecasting with neural networks at uber', in *International Conference on Machine Learning*, Vol. 34, pp.1–5.

Last, M., Klein, Y. and Kandel, A. (2001) 'Knowledge discovery in time series databases', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, February, Vol. 31, No. 1, pp.160–9, https://doi.org/10.1109/3477.907576.

Lawhern, V., Hairston, W.D. and Robbins, K. (2013) 'Detect: A MATLAB toolbox for event detection and identification in time series, with applications to artifact detection in EEG signals', *PloS One*, 24 April, Vol. 8, No. 4, p.e62944, https://doi.org/10.1371 /journal.pone.0062944.

Lee, M. and Kim, H.J. (2018) 'Construction of event networks from large news data using text mining techniques', *Journal of Intelligence and Information Systems*, Vol. 24, No. 1, pp.183–203.

Li, W., Gao, X., Hao, Z. and Sun, R. (2022) 'Using deep learning for precipitation forecasting based on spatio-temporal information: a case study', *Climate Dynamics*, Vol. 58, No. 1, pp.443–457.

Li, X. and Lin, J. (2017) 'Linear time complexity time series classification with bag-of-pattern-features', in *2017 IEEE International Conference on Data Mining (ICDM)*, pp.277–286, IEEE.

Ling, X. and Weld, D.S. (2010) 'Temporal information extraction', in *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Liu, X., Zheng, L., Zhang, W., Zhou, J., Cao, S. and Yu, S. (2022) 'An evolutive frequent pattern tree-based incremental knowledge discovery algorithm', *ACM Transactions on Management Information Systems (TMIS)*, Vol. 13, No. 3, pp.1–20.

Madan, R. and Sarathi Mangipudi, P. (2018) 'Predicting computer network traffic: a time series forecasting approach using DWT, ARIMA and RNN', in *IEEE Eleventh International Conference on Contemporary Computing (IC3)*, pp.1–5.

Mehndiratta, P. and Soni, D. (2019) 'Identification of sarcasm in textual data: a comparative study', *Journal of Data and Information Science*, Vol. 4, No. 4, pp.56–83.

Morchen, F. and Ultsch, A. (2005) 'Optimizing time series discretization for knowledge discovery', in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 21 August, pp.660–665, https://doi.org/10.1145/1081870.1081953.

Music Dataset MER 500 (2020) [online] https://www.kaggle.com/datasets/makvel/mer500 (accessed 25 June 2022).

Ohga, Y. and Seki, H. (1993) 'Abnormal event identification in nuclear power plants using a neural network and knowledge processing', *Nuclear Technology*, 1 February, Vol. 101, No. 2, pp.159–67, https://doi.org/10.13182/NT93-A34777.

Pandey, S., Srivastava, A. and Amidan, B. (2020) 'A real time event detection, classification and localization using synchrophasor data', *IEEE Transactions on Power Systems*, Vol. 35, No. 6, pp.4421-4431, DOI: 10.1109/TPWRS.2020.2986019.

Pekkola, S., Niemi, P. and Ukko, J. (2013) 'Building understanding of the development of performance management for collaborative networks with a knowledge maturity model', *International Journal of Networking and Virtual Organisations*, 1 January, Vol. 12, No. 3, pp.179–200, https://doi.org/10.1108/IJOPM-10-2013-0469.

Pretet, L., Richard, G., Souchier, C. and Peeters, G. (2022) 'Video-to-music recommendation using temporal alignment of segments', *IEEE Transactions on Multimedia*, DOI: 10.1109/TMM. 2022.3152598.

Qi, H., Xiao, S., Shi, R., Ward, M.P., Chen, Y., Tu, W. and Zhang, Z. (2020) 'COVID-19 transmission in Mainland China is associated with temperature and humidity: a time-series analysis', *Science of the Total Environment*, August, Vol. 728, p.138778, DOI: 10.1016/ j.scitotenv.2020.138778.

Qin, M., Li, Z. and Du, Z. (2017) 'Red tide time series forecasting by combining ARIMA and deep belief network', *Knowledge-Based Systems*, June, Vol. 125, pp.39–52.

Radaideh, M.I., Pigg, C., Kozlowski, T., Deng, Y. and Qu, A. (2020) 'Neural-based time series forecasting of loss of coolant accidents in nuclear power plants', *Expert Systems with Applications*, p.113699, https://doi.org/10.1016/j.eswa.2020.113699.

Ramirez-Gallego, S., Krawczyk, B., García, S., Woźniak, M. and Herrera, F. (2017) 'A survey on data preprocessing for data stream mining: Current status and future directions', *Neurocomputing*, 24 May, Vol. 239, pp.39–57, https://doi.org/10.1016/j.neucom.2017.01.078.

Rao, P., Ross, J.C., Ganguli, K.K., Pandit, V., Ishwar, V., Bellur, A. and Murthy, H.A. (2014) 'Classification of melodic motifs in raga music with time-series matching', *Journal of New Music Research*, Vol. 43, No. 1, pp.115–131.

Selim, A., Abdel-Akher, M., Kamel, S. and Aly, M.M. (2018) 'A developed approach based on lagrange linear prediction for time-series power-flow simulation', *Electric Power Components and Systems*, Vol. 46, Nos. 11–12, pp.1312–1320.

Shahar, Y., Stein, A. and Musen, M.A. (1996) 'Knowledge acquisition for temporal abstraction', in *Proceedings of the AMIA Annual Fall Symposium*, pp.204–208 [online] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233051/ (accessed 15 June 2022).

Shannon, C.E. (1948) 'A mathematical theory of communication', *The Bell System Technical Journal*, Vol. 27, No. 3, pp.379–423.

Sharma, A.K., Aggarwal, G., Bhardwaj, S., Chakrabarti, P., Chakrabarti, T., Abawajy, J.H., ... and Mahdin, H. (2021) 'Classification of Indian classical music with time-series matching deep learning approach', *IEEE Access*, July, Vol. 9, pp.102041–102052, DOI: 10.1109/ACCESS.2021.3093911.

Si, Y. (2021) 'Construction of music storage and extraction system based on time series analysis', in *2021 4th International Conference on Information Systems and Computer Aided Education*, September, pp.506–510.

Singh, G., Hawkins, L. and Whymark, G. (2007) 'An integrated model of collaborative knowledge building', *Interdisciplinary Journal of E-Learning and Learning Objects*, 1 January, Vol. 3, No. 1, pp.85–105 [online] https://www.learntechlib.org/p/44800/ (accessed 15 June 2022).

Singh, T., Kumari, M. and Gupta, D.S. (2022) 'Real-time event detection and classification in social text steam using embedding', *Cluster Computing*, pp.1–19, https://doi.org/10.1007/s10586-022-03610-6.

Sirichanya, C. and Kraisak, K. (2021) 'Semantic data mining in the information age: a systematic review.', *International Journal of Intelligent Systems*, Vol. 36, No. 8, pp.3880–3916.

Sonawane, S.S. and Kulkarni, P.A. (2016) 'Context-based co-reference resolution for text document using graph model (cont-graph)', *International Journal of Knowledge Engineering and Data Mining*, Vol. 4, No. 1, pp.1–17.

Souto, G. and Liebig, T. (2016) 'On event detection from spatial time series for urban traffic applications', in *Solving Large Scale Learning Tasks, Challenges and Algorithms*, pp.221–233, Springer, Cham.

Thiesen, S., Darscheid, P. and Ehret, U. (2019) 'Identifying rainfall-runoff events in discharge time series: a data-driven method based on information theory', *Hydrology & Earth System Sciences*, Vol. 23, No. 2, pp.1015–1034.

Tian, D., Li, M., Shi, J., Shen, Y. and Han, S. (2021) 'On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach', *Advanced Engineering Informatics*, August, Vol. 49, p.101355, https://doi.org/10.1016/j.aei.2021.101355.

Tsien, C.L. (2000) 'Event discovery in medical time-series data', in *Proceedings of the AMIA Symposium*, p.858, American Medical Informatics Association [online] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243881/ (accessed 15 June 2022).

Ultsch, A. (1999) 'Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series', in *Kohonen Maps*, 1 January, pp.33–45, Elsevier science BV, https://doi.org/10.1016/B978-044450270-4/50003-6.

Usai, A., Pironti, M., Mital, M. and Mejri, C.A. (2018) 'Knowledge discovery out of text data: a systematic review via text mining', *Journal of Knowledge Management*, Vol. 22, No. 7, pp.1471–1488, https://doi.org/10.1108/JKM-11-2017-0517.

Velankar, M. and Kulkarni, P. (2018) 'Unified algorithm for melodic music similarity and retrieval in query by humming', in *Intelligent Computing and Information and Communication*, pp.373–381, Springer, Singapore.

Wang, D., Zhang, X., Wan, Y., Yu, D., Xu, G. and Deng, S. (2021) 'Modeling sequential listening behaviors with attentive temporal point process for next and next new music recommendation', *IEEE Transactions on Multimedia*, DOI: 10.1109/TMM.2021.3114545.

Wu, Y., Lin, Y., Zhou, Z., Bolton, D.C., Liu, J. and Johnson, P. (2018) 'DeepDetect: a cascaded region-based densely connected network for seismic event detection', *IEEE Transactions on Geoscience and Remote Sensing*, 2 August, Vol. 57, No. 1, pp.62–75, https://doi.org/10.1109/TGRS.2018.2852302.

Yamanishi, K. and Takeuchi, J.I. (2002) 'A unifying framework for detecting outliers and change points from non-stationary time series data', in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 23 July, pp.676–681, https://doi.org/10.1145/775047.775148.

Zhou, Y., Arghandeh, R. and Spanos, C.J. (2017) 'Partial knowledge data-driven event detection for power distribution networks', *IEEE Transactions on Smart Grid*, Vol. 9, No. 5, pp.5152–5162.