International Journal for Modern Trends in Science and Technology, 8(12): 11-14, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0812003
Available online at: http://www.ijmtst.com/vol8issue12.html

An Online Approach Towards Web Content Extraction using Machine Learning Algorithms

Dr. Kishor M. Dhole[1] | Dr. Rajesh K. Parate[2]

[1]Assistant Professor, Department of Computer Science, Seth Kesarimal Porwal College of Arts and Science and Commerce, Kamptee, Nagpur (India)-441001

[2]Assistant Professor, Department of Electronics, Seth Kesarimal Porwal College of Arts and Science and Commerce, Kamptee, Nagpur (India)-44100

To Cite this Article

Dr. Kishor M. Dhole and Dr. Rajesh K. Parate. An Online Approach Towards Web Content Extraction using Machine Learning Algorithms. International Journal for Modern Trends in Science and Technology 2022, 8(12), pp. 11-14. https://doi.org/10.46501/IJMTST0812003

Article Info

Received: 16 November 2022; Accepted: 27 November 2022; Published: 02 December 2022.

ABSTRACT

The Internet based World Wide Web has seen gigantic development as of last decade. With the huge measure of data on the various websites, webpages have been the likely wellspring of data recovery and information mining innovation, for example, business web search tools, web mining applications. Web pages contain a few things that can't be delegated the instructive substance, e.g., search and sifting board, route connections, promotions, etc called as strident parts. Most clients and end-clients look for the useful substance, and generally don't need the non-educational substance.

A tool that helps an end-user client or application to look and handle data consequently, should isolate the "essential or educational substance segments" from the other substance areas. The substance extraction issue has been a subject of concentrate since the extension of the Internet. Its will probably isolate the primary substance of a page, like the text of a report, from the uproarious substance, for example, ads and route joins. Most content extraction approaches work at a block level; that is, the website page is portioned into blocks and afterward every one of these still up in the appearance to be essential for the principal content or the noisy satisfied of the webpage. The extricated primary substance is summed up into tabular format for the better understanding of identified information from the source.

KEYWORDS: WebPages, WebCrawler tools, Support Vector Machines (SVM), Clustering, K-Means algorithms.

1. INTRODUCTION

The World Wide Web online pages of any source of information are known as web documents. These WebPages are the sources for a wide range of informational categories [1].

These include, to name a few, news,encyclopaedia articles, forum debates, and advertisementsforproducts.Each type of information can have a variety of media types, including textual, graphical, and visual[2]. This large volume of data is used by both automated crawlers who scour the web for different purposes, including web mining or web indexingmethods[3][4][5].Theregular web users from all over the world.However, a single webpage typicallycomprises of various "pieces," which throughout this research paper will be referredto as the webpage's contents[5]. Th

ere is just one form of content, which will be referred to as prior content of information extraction environment[6][7].A variety of applications targeted at comprehending the online are made possible by a variety of technologies, including web content extraction. The specific type of text contents, which will be refer as the key content of the webpage via online mode of information retrieval system. This makes the webpage a useful source of information for extraction of information from the online source like as website [8]. Other contents include advertisements, navigation buttons, page settings, and legal notices; these contents will be collectively referred to as the strident content of the webpage [9]. The process of identifying the main content of a web page via online is called online content extraction, or more briefly content extraction. This needs a scientific approach to identify the content information in a very short time. Thus it needs to study various algorithms using machine learning approach [10].

LITERATURE REVIEW

Following machine learning algorithms techniques has been reviewed for online web content extractions

TIME SERIES ANALYSIS

From a statistical perspective, the detected values frequently represent the observations of a random sample of independent random variables. The type of dependence between the sequence's components forms the basis for the time series analysis. With this method, many events are no longer treated as being haphazardly distributed throughout time around a rather steady average.The concept of memory, persistence, or hysteresis must be at the core of the study of time series.[11],[12].The stochastic process is the method we employ to model the observed time series. A stochastic process is conceptualized intuitively as an endlessly long chain of random variables or an infinitely large random vector [13][14]. The degree of link between the random variables that make up a stochastic process, which determines its memory, determines how a sample of consecutive observations throughout time should be viewed of rather than as realizations of t separate random variables [15][16].The deterministic and random disturbance components of the process are assumed to exist in the classical approach to time series [17][18].

TIME SERIES DATA SETS

To apply machine learning algorithm, perhaps it has already been done away with. The stochastic component, which is thought to be a process with regulated components, is the main focus[19]. Due to the limited number of accessible observations, the stochastic process can never be precisely recognised; nonetheless, we can try to recreate it using a model [20].However, a model that simulates the datagenerating process must have a few features in order to be identifiable with reasonable accuracy. It must be possible to invert the process, it must be Gaussian, and it must be possible to invert the process from the time series[21].

TIME SERIES DATA ANALYSIS USING MACHINE LEARNING ALGORITHMS: AI based machine learning calculations independently foster their insight thanks to the information designs got, without the need to have explicit beginning contributions from the engineer [22]. In these models, the machine can lay out without help from anyone else the examples to follow to get the ideal outcome, consequently, the genuine element that recognizes man-made brainpower is independence. In the growing experience that recognizes these calculations, the framework gets a bunch of information essential for preparing, assessing the connections between the information and yield information: these connections address the boundaries of the model assessed by the framework [23].

THE PROBLEM STATEMENT

Web extraction has been extensively investigated, although they frequently concentrate on extracting structured data from webpages that include several instances of the same structured data, such as product catalogues[24] .This project seeks to extract less organised online material from noisy webpages, such as news pieces that only appear once. Our strategy employs a combination of visually appealing and linguistically neutral characteristics. Additionally, a pipeline is created to automatically label data points by clustering, where each cluster is graded according to how well it matches the webpage description that was taken from the best clusters data points and meta tags are chosen as a good examples [25].

PROPOSED METHODOLOGY

The proposed method focuses on web pages with unstructured text as the primary source of information.The entire webpage is subject to the information extraction technique, and only the key content blocks of the web pages are really searched for information. Information on the user species is necessary for the system.Starting from one o

r more seed URLs, web crawlers download all associated pages, extract the hyperlink URLs from those sites.

Following steps has been suggested by researcher for successful implementation of web content information extraction mechanism in online mode.

i. Selection of URL Webpages
ii. Choosing webcrawler for identification and analyzing of text
iii. Clustering of text using webcrawler
iv. Analysis of clustered content information as relevant data and irrelevant data
v. Apply algorithmic approach for classification of relevant data content and irrelevant data content
vi. Finalization and interpretation of summarized text from resource website.

WEB CONTENTS SYSTEM FLOW

System flow for web content of information extraction is as follows:

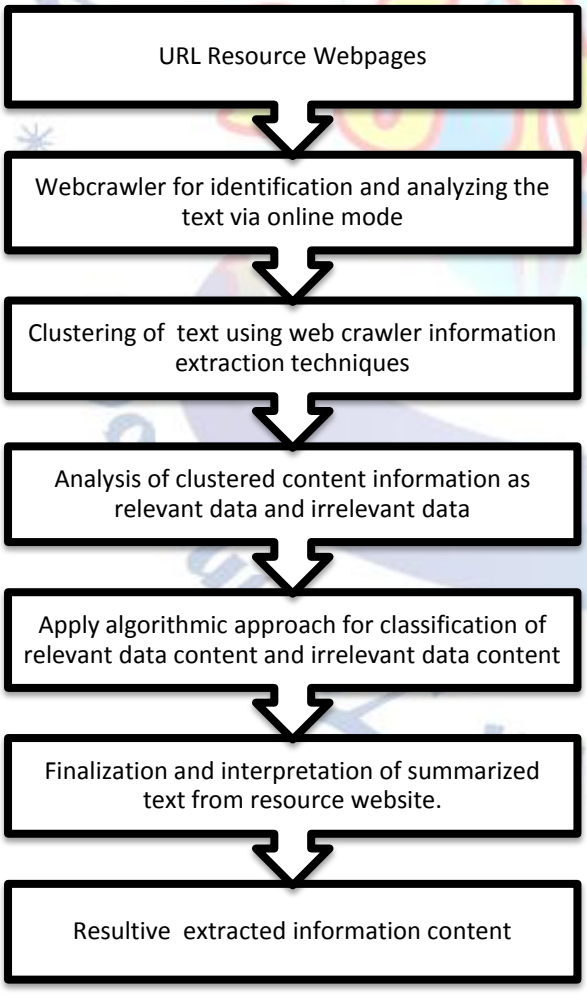

Figure 1: System Flow of Web Content Extraction

WORKING MODEL STAGES

The proposed working model for the system has a main website where users enter the URL of the webpage whose data has to be crawled and it operates by using following stages to gain accurate content from the web is as follows:

Enter the site's URL

A crawler will investigate the site and accumulate text information.

Use bunching to isolate the recovered information into groups like text, joins, and so on.

Assigning a 1 to information that is associated with that page and a 0 to information that is irrelevant.

Use SVM or any Machine Learning algorithm to separate information into content and non-content classes.

This method will kill copy information, increases, and clamor.

The site's text outline will be the last result.

CONCLUSION

The above system flows collects information, identifies web content models using algorithmic approach using machine learning and classify it according to clustering mechanism, trains support vector classifier, and assesses learned model in a android manner. The learning and trained algorithm could accomplish wonderful marking of online content based information when prepared on a solitary site. The researcher incorporated these modification for achieving better results as: The rundown is acted in even configuration. Rather than showing connections of pictures, the portrayal of the pictures shown in linkages.

The links to the online web content information can be made dynamic for the reference of human clients.

The framework can be scaled to deal with a site all in all, and, surprisingly, further for numerous sites. Nonetheless, the memory required and the expense related for a similar will be exceptionally high. For business execution, this choice merits attempting.

Again from human client's experience concern, the clustering of text, pictures and connections can act naturally independently showed.

A nonstop admittance to all pictures in a successive way can likewise be publicized in a proper sequence for the content identification and extraction is more fruitful.

FUTURE IMPLICATIONS

As per above algorithmic approach researchers may choose other machine learning algorithms for gaining good results. These algorithms are Artificial Neural Network-Based Methods, Time Series Clustering Methods, Convolution Neural Network for Time Series Data, Recurrent Neural Network, Auto encoders Algorithms in Time Series Data Processing, Automated Features Extraction from Time Series Data.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] C. Kohlsch¨utter, P. Fankhauser, and W. Nejdl Boilerplate detection using shallow text features. In Proceedings of WSDM '10, pages 441–450. ACM, 2010.

[2] J. Pasternack and D. Roth. Extracting article text from the web with maximum subsequence segmentation. In Proceedings of WWW '09, pages 971–980. ACM, 2009.

[3] F. Sun, D. Song, and L. Liao. Dom based content extraction via text density. In SIGIR, volume 11, pages245– 254, 2011.

[4] T. Weninger, W. H. Hsu, and J. Han. CETR: content extraction via tag ratios. In Proceedings of WWW '10,pages 971–980 .ACM,2010.

[5] Wei, W.W. Time series analysis. In The Oxford Handbook of Quantitative Methods in Psychology; Oxford University Press: New York, NY, USA, 2006; Volume 2.

[6] Lütkepohl, H. New Introduction to Multiple Time Series Analysis; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005.

[7] Chatfield, C.; Xing, H. The Analysis of Time Series: An Introduction with R; CRC Press: Boca Raton, FL, USA, 2019.

[8] Hamilton, J.D. Time Series Analysis; Princeton University Press: Princeton, NJ, USA, 2020.

[9] Brillinger, D.R. Time Series: Data Analysis and Theory; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2001.

[10] Granger, C.W.J.; Newbold, P. Forecasting Economic Time Series; Academic Press: Cambridge, MA, USA, 2014. Cryer, J.D. Time Series Analysis; Duxbury Press: Boston, MA, USA, 1986; Volume 286.

[11] Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. Time Series Analysis: Forecasting and Control; John Wiley & Sons: Hoboken, NJ, USA, 2015.

[12] Madsen, H. Time Series Analysis; CRC Press: Boca Raton, FL, USA, 2007.

[13] Fuller, W.A. Introduction to Statistical Time Series; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 428.

[14] Tsay, R.S. Analysis of Financial Time Series; John Wiley & Sons: Hoboken, NJ, USA, 2005; Volume 54.

[15] Harvey, A.C. Forecasting, Structural Time Series Models and the Kalman Filter; Cambridge University Press: Cambridge, UK, 1990.

[16] Kantz, H.; Schreiber, T. Nonlinear Time Series Analysis; Cambridge University Press: Cambridge, UK, 2004; Volume 7.

[17] Shumway, R.H.; Stoffer, D.S.; Stoffer, D.S. Time Series Analysis and Its Applications; Springer: New York, NY, USA, 2000; Volume 3.

[18] Fahrmeir, L.; Tutz, G.; Hennevogl, W.; Salem, E. Multivariate Statistical Modelling Based on Generalized Linear Models; Springer: New York, NY, USA, 1990; Volume 425.

[19] Koustas, Z.; Veloce, W. Unemployment hysteresis in Canada: An approach based on long-memory time series models. Appl. Econ. 1996, 28, 823–831.

[20] Teyssière, G.; Kirman, A.P. (Eds.) Long Memory in Economics; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.

[21] Gradišek, J.; Siegert, S.; Friedrich, R.; Grabec, I. Analysis of time series from stochastic processes. Phys. Rev. E 2000, 62, 3146. [CrossRef] [PubMed]

[22] Grenander, U.; Rosenblatt, M. Statistical spectral analysis of time series arising from stationary stochastic processes. Ann. Math. Stat. 1953, 24, 537–558.

[23] Alessio, E.; Carbone, A.; Castelli, G.; Frappietro, V. Second-order moving average and scaling of stochastic time series. Eur. Phys. J. B Condens. Matter Complex Syst. 2002, 27, 197–200.

[24] Klöckl, B.; Papaefthymiou, G. Multivariate time series models for studies on stochastic generators in power systems. Electr. Power Syst. Res. 2010, 80, 265–276.

[25] Harvey, A.; Ruiz, E.; Sentana, E. Unobserved component time series models with ARCH disturbances. J. Econom. 1992, 52, 129–157.

[26] Nelson, C.R.; Plosser, C.R. Trends and random walks in macroeconomic time series: Some evidence and implications. J. Monet. Econ. 1982, 10, 139–162.

[27] Hagan, M.T.; Behr, S.M. The time series approach to short term load forecasting. IEEE Trans. Power Syst. 1987, 2, 785–791.

[28] Ciaburro, G. Sound event detection in underground parking garage using convolutional neural network. Big Data Cogn. Comput. 2020, 4, 20.

[29] Mohri, M.; Rostamizadeh, A.; Talwalkar, A. Foundations of Machine Learning; MIT Press: Cambridge, MA, USA, 2018.