International Journal of Engineering & Technology, 7 (4.37) (2018) 168-172

International Journal of Engineering & Technology

Website: www.sciencepubco.com/Index.Php/IJET

Research Paper

Big Data Analysis of Web Data Extraction

Nadia Ibrahim[1], Alaa Hassan[2] , Marwah Nihad[3]

[1,2,3]Kirkuk University, Kirkuk, Iraq
**Corresponding author E-mail: nadia.ibra@uokirkuk.edu.iq*

Abstract

In this study, the large data extraction techniques; include detection of patterns and secret relationships between factors numbering and bring in the required information. Rapid analysis of massive data can lead to innovation and concepts of the theoretical value. Compared with results from mining between traditional data sets and the vast amount of large heterogeneous data interdependent it has the ability expand the knowledge and ideas about the target domain. We studied in this research data mining on the Internet. The various networks that are used to extract data onto different locations complex may appear sometimes and has been used to extract information on the web technology to extract and data analysis (Marwah et al., 2016). In this research, we extracted the information on large quantities of the web pages and examined the pages of the site using Java code, and we added the extracted information on a special database for the web page. We used the data network function to get accurate results of evaluating and categorizing the data pages found, which identifies the trusted web or risky web pages, and imported the data onto a CSV extension. Consequently, examine and categorize these data using WEKA to obtain accurate results. We concluded from the results that the applied data mining algorithms are better than other techniques in classification and extraction of data and high performance.

Keywords: *Web data extracting, classification, data mining algorithms, WEKA.*

1. Introduction

At present with the great development of the information technology revolution, and facilitate the millions of people through a huge database of increase issue and continue to practice sensors and a variety of digital devices has come, resulting in so-called "big data". Big data is word of relating to huge sizes of difficult datasets (Nelofar,2017). Big data is an abstract idea. It also has some other features, apart from the masses of data, which specifies the variance between large data or very large data. It needs appropriate handling power and great abilities for exploration (Boyd and Crawford, 2011). Big data analysis as an important activity for many organizations emerged. This is to simplify large data analysis of the frame and the implementation environment, such as Hadoop systems and parallel, like the beehive. Data mining method shows active role in the examination of data (Jharna *et al*., 2016). Large additions require high associated data capture and analysis, as well as the results predictive reports. With large data, and better organization of information technology in all parts of specific and a potential opportunity rather than just a set of common services that serve both traditional and uses the latest. This phenomenon is confirmed that the massive amounts of data generated and constantly increase ever and the unprecedented levels, found improvement of existing algorithms and techniques and techniques through the training of parallel computing architectures (cloud platforms in our minds). And also you must deal with the lack of homogeneity when large data mining and privacy-scale and speed, confidence and accuracy, and that the current mining algorithms and methods are capable to interact and the need to design and implement a parallel machine learning and large data extraction range using algorithms has increased remarkably, that accompany the emergence of a powerful parallel processing platforms and data on a very large scale, for example, Hadoop map reduce. Big data extraction must deal with semi-structured and unmatched data. Simple example is mentioned by growing the knowledge to the online marketplace, like eBay. Currently a dataset is a rich network of data, which is composed of three kinds of objects: sellers, items, and buyers (where there is a large data extraction is complex). For example, there may be a correlation between large data widely, between the items and buyers, items of goods, sellers and items, and between buyers and sellers. This large data have several forms of objects and relationships. Usually hidden relationships in the large amount of data to be interesting, and the extraction and exploration of these data reveals patterns and relationships between them, and the results of the query for this data will help make future opinions in the world, as well as make predictions value. And a wide field of applications like Areas of Medicine, business, engineering and science applies data mining. This has in turn led to a lot of a lot of real companies - where the benefit to each of the service providers as well as consumers of services useful services. Overcome this challenge, the viability of large data, made many tries to take benefit of huge parallel processing structures. And the first attempted was made through Google. Google has produced a programming model called Mabredeos. In addition to the GFs (Google File System) and a distributed file system for large data route can be divided on thousands of nodes in a cluster. And then, Yahoo has created and several main companies Apache open source version of the plane Mabredeos Google structure called Hadoop map reduce name. And it utilizes Hadoop Distributed File System (HDFS) - an open source version of GFS in Google. Map minimize framework that permit users to specify the two functions, map and minimize, to handle a large of mount of data entered in parallel. The data is analyzed to extract these huge data. With information technology and easy access to a large volume of information creates anxiety of large databases on the Internet, from where of the existence of these random sources. This paper talks about extensive data, as

well as talking about areas that cannot be extracted from various kinds of data sources. Large information everywhere and this in turn will increase the necessary tools and sophisticated and smart to check the data and information of mine and knowledge of them. For example of the traditional statistical dealing with this high rate and employ advanced methods to provide this information and analysis. To mine the chief content of the website, data mining methods require being applied (Neha and Saba, 2011). A goal of data mining technique is to extract familiarity from large information based on procedures, which are assembled to extract information and from various fields, like mathematics, statistics and logic, artificial intelligence and expert systems. Data mining is a developed exploration of a big size of data to find novel information in the summary of designs (Chandaka *et al*., 2018). The importance of methods of extracting Web data depends on the details of the joint between the huge information. Data mining on the Internet to collect this data through the banned human power, which is one of the smart systems allow and non-traditional science. It has found many ways to capture and get the data from the Internet to solve some problems. A big volume of data indicates to the huge data problem (Rajkumar and Usha, 2016). Data mining to find a large data include steps computational algorithms complex. Data mining is a programmed method applied to mine beneficial information from big and compound data sets (Manisha *et al*., 2015). Network extract may be the basis for the exploration of large data. Data mining can collect the entire data basis and if this process needs data that can be obtained from the Internet and then Internet extract probably one of the methods to find this information. Information acquisition is one of the affirmative outcomes of the investigation into the large amounts of information, as it turns information that has been collected and that is incomprehensible to the value of interest and can be applied in the knowledge of later information. There are different benefits in the field of data mining in an attempt to manage the growing data extraction with the cognitive patterns of algorithms, as well as the development of scalable. With this technology available, it has evolved and expanded software and mining algorithms is very large. So the major aim of data mining is to draw interest and knowledge of the content of large data, including Internet data. The method works to extract information from web pages as well as the mining Web data. Web mining is the practice of data mining methods to mine valuable information from web data (Pranit and Sheetal, 2018). The purpose of Web Mining is to learn and recover beneficial and interesting outlines from precise huge web dataset (Lourdu *et al*., 2016). It utilizes data mining approaches allowed to obtain important data and useful information on the Internet. WWW or big data containing of unlike information to satisfy our desires based on the DM procedures and web systems and the detection of knowledge of the actuality of all the big data existing on the internet.

2. Web Content Extractions

This type of study on the Internet to search for knowledge of the sites pages on Internet content, and comprises in this pattern on the sites and in accordance with their themes, its content is categorized and the assessment and extraction of content and analyze their data. This type contain knowledge detection from the actually of the check of remarks and response from the beneficiaries and readers of the importance displays that can be invested in several features of knowledge, it should be distinguished that this does not impact to data mining as it is not existing in the ability tables of database add notes or responses performance on the content. Content mine from webpage is an important stage for information achievement (Gunasundari and Karthikeyan, 2012). Exploration data has to transform a variety of web pages. And it is linked to the work of analysis and collection of information for the web exactly for data extraction. Idea of pattern and relational data mining utilized to conclude the relationship procedures in the text.

2.1. Data Mining Algorithms

There are some approaches in order to practice this problem: clustering, association, classification.

1. Associations: It intended to detect correlations between groups of elements. Association and resulting repeated item groups. Association and relationship is commonly to discover many item set results between huge data sets (Bharati, 2010). Associations Granted maintenance and guarantee well results in many areas as big data. Mining rule Assembly has great sequences of navigation training the application of such a site. It is easy to train and exercise.

2. Classification: It's one of the greatest normally data extraction process technology. Classification method is mastered to deal with an extensive difference of the big data and developing in regard. The classification comprises estimated outcome is guaranteed based on the expected input. In order to analyze the result, the system is trained digit constant preparation of properties held and the result is obvious. Classification of the normal examines of information, training and constructs a model for each sets reliant on the structures in the data, like classified beam of automated support, decision trees, menus, based on, for example, multilayered receptors algorithm, logistic regression, and Obaiz networks. Classification experiment data are practical to evaluation the precision of the classification procedures (Bhu *et al*., 2014). Among different forms of awareness of the present recording in the classification it is regularly used instructions of arithmetic notes to study.

3. Clustering: Clustering is important assignment in data exploration and data mining uses (Chitra1 and Maheswari, 2017). It is to build big data clusters process, as well as collect similar data set with each other, meaning that objects that are within the cluster be identical.

2.2. Decision Tree Algorithm

The decision tree algorithms are typically utilized big data classification. It is illustrative drawing and designated tree chart. Trees decision procedure are securities that manage structures by classified according to standard feature. Each node is characterized by the feature of instance of a classification. All branches reviews the value of the node can agree. The first structures are organized in the source node (root) and categorized based on their features. Just can modification the decision of trees calculated to IF-THEN commands. This is the algorithm used for the classification of features. Everywhere it is input to the normal classification algorithm for big data output and more than that you can establish new data not earlier practical as a kind of account of this original information. Decision tree classifiers find superior correctness while matched with other classification techniques (Bhu *et al*., 2012). This workbook can be enhanced in the structure method of a tree, motivation the decision tree in the order of the rules of so-called rules of decision-making. This is the technique utilized divide and conquer standard the knowledge of dividing the issue into parts and solve them independently and currently collected outcome. Decision tree can practice individually constant and definite data (Himani and Sunil, 2015). Decision tree was recognized depended on the greatest feature of the property choice and exercise can be set so that the division of the depth of the tree at the equivalent time at smallest categorized data accurately. In conclusion, decisions that effectively constructed tree. Decision tree competition the fast models of other classification. It is simple to character out the classification processes

2.3. Analysis Web

Extraction of useful information on the Internet is not a simple process. Web where he plays a key role in the exploration and also provides tools that help extract knowledge and interest of the suggestion on the Internet Web indicates that it is the main sources of

information foundations (big data) in the world. We be certain of that any subject and in any field is now located on the web page. Data on the internet is unlike types and shapes like pictures, text, and video. In this research, we used data mining the Web for the discovery of protected pages techniques. Whole data is important difficulties for precision big data. Assume well information is accessed; the next step is to recognize the best appropriate big data for the test process. Display information equivalent association existing for discovery the best results.

2.4. WEKA Data Mining System

WEKA is Waika to area for Knowledge Examination. WEKA's a data mining machine learning appliance progressive via, Computer Science Department, New Zealand, and University of Waikato. It is free data mining system existing. Weka is a set of machine learning systems for data mining purposes.
The developments can equally be utilized to a dataset or practical from your isolated Java code. Weka contains implements for information pre-management, classification, association rules, visualization, regression, and clustering.

3. Methodology

1- Research Problem

One challenge is to manually fetch the required data from the web content. Data collection costs increase as more data is collected from each site, making it difficult to extract information. The Internet as well as data growth has made it more difficult for us to extract big data in a timely manner, so it is difficult to collect the required web pages and extract useful data for analysis. The study shows the discovery and extraction of big data from web pages and analyzes and verifies data. Using data mining techniques to extracting web content and detect unreliable sites and trusted sites on the Internet.

2- Basic Models

Dataset: A static of data elements; the dataset is a specific easy knowledge of machine learning. The dataset is almost identical to two-dimensional worksheet or database table. In WEKA, it is carried out through the Instances.

Instance: contains of features (Attributes).

Item set: A set of items that appear regularly grouped in a transactions dataset.

Attributes:

- Nominal: one of a predefined list of values.
- Numeric: A real or integer number.
- String: Enclosed in "double quotes".
- Date
- Relational.

Trust: The trust is well-clear via a practiced probability Assurance(X=>Y) =Delivery (XUY)/ Delivery (X) =P(Y/X).
Lift: It is the relation of the probability that L and R occur grouped to the two different potentials for L and R. This study explains the utilized of java to scan HTML pages and to obtain page content, then mined the web pages content into "BigData-webhacker.arff" file. The case big dataset practical for this case is based on the "BigData-webhacker.arff" via applying a **decision tree** process through WEKA package. This study supposes that suitable big data preprocessing has been done. The URLs and the information extracted from these URLs are displayed in (Fig. 1) and (Fig. 2):

Fig. 1: Analysis large numbers of URL (HTML pages) to retrieve page content and extracted big data using java

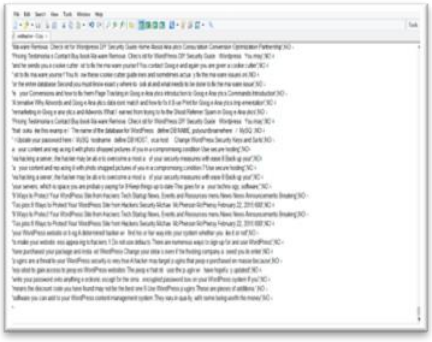

Fig. 2: "BigData-webhacker.arff" mined from webpages

3.1. Related Work and Results

The proposal contains:

- The data set name – "BigData-webhacker.arff".
- Size of data set = 8.47 GB (9,096,733,606 bytes), 71,091,606 lines.
- Number of cases in the relation (row) = 6640 URLs.
- Number of features in the table = 3.

Attribute description:

Id: definition number

Words: string

Hackers :{ yes, no}

- **Upload dataset:**

- Upload the big data in Weka (press on Preprocess and then on Open file... "BigData-webhacker.arff").

- **Decision tree procedure to the "BigData-webhacker.arff"**

- As soon as learning the data executed the delivered procedures.

- Examine the dataset through C4.5 process involving J48, WEKA's service of decision tree practiced. The case data used in this run is the web data from the "BigData-webhacker.arff".

- Once you have your data employed, completely the tabs are available to you. Press on 'Classify' tab
- Press on 'Choose' in the 'Classifier' and select C4.5

- Classifier WEKA - Classifiers - Trees - J48.

- In this application, you will evaluation classifier based on how well it supposes.
- After training is perfect as displayed in (Fig. 3) and (Fig. 4):

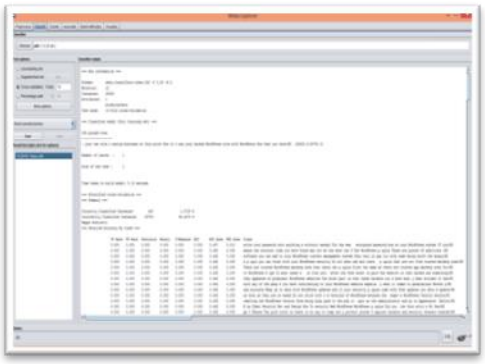

Fig. 3: Training big dataset using - Trees - J48

Fig. 4: Classifier Visualize of big data

3.2. The 'Classifier' outcome range calls the outcomes of training and testing such as following:

The static of totals is resulting from the execution data. In this state, 98% of 29253 training cases were classified correctly. This shows that the outcomes from training data are not guaranteed complemented with what might be developed from the permitted test set from the corresponding basis. Also to organization incorrectness, the valuation outcome measurements resulting from the session potentials assigned through the tree. Further exactly, it results means result fault (0.1) of the probability estimates, the root mean squared fault (0.3) is the square root of the quadratic loss. The mean absolute error calculated in a same manner via applying the total in its place of squared change. The reason that the errors are not 0 or 1 is because not fully training instances are classified correctly.

3.3. Next performing the clustering algorithm and finds the following results:

Fig. 5: Training big data using clustering algorithm

3.4. Visualize task clustering is as following:

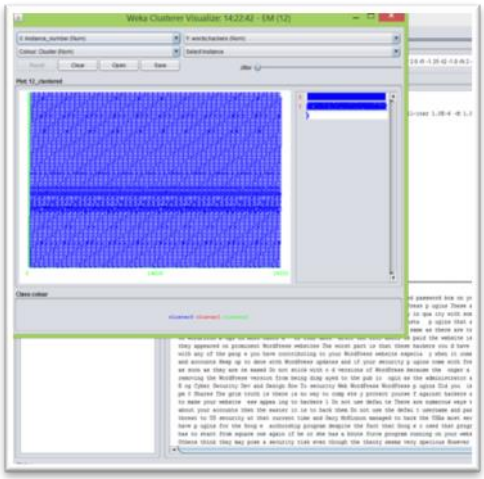

Fig. 6: Visualize data clustering

4. Conclusions

In this study, analyze and extract large data for many websites, as well as evaluation and categorizing web pages into confident Web pages and non-confident Web pages. We conclude that there is a big of data on the Internet, where mining on the Internet uses various techniques to extract data to discover useful knowledge of web content and page. After finding the data, this large data will be tested using data extraction methods, where the pages are evaluated for accurate results in the usage classification and algorithm aggregation. This evaluation relies on data from these pages using the decision tree on WEKA. We have found that the procedures used by others are improved in performance. In the future, we are expanding a software package to fully extract big data from websites. We trust in the future that we are using the largest procedures to get the best results in the extraction of web information.

References

[1] Abdullah, Marwah N., Alaa Hassan, and Nadia Naef. , 2016, Knowledge-Based Analysis of Web Data Extraction, Proceedings of the Fifth International Conference on Informatics and Applications, Takamatsu, Japan, ISBN: 978-1-941968-41-3 SDIWC 26.

[2] Bharati M., 2010, Data Mining Techniques and Applications, Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305.

[3] Bhu L., Arundathi, and Jagadeesh, 2014, Data Mining: A prediction for Student"s Performance Using Decision Tree ID3 Method, International Journal of Scientific & Engineering Research, Volume 5, Issue 7, July-2014 1329 ISSN 2229-5518.

[4] Boyd D., and Crawford K., 2011, Six provocations for big data. In A decade in internet time: Symposium on the dynamics of the internet and society ", Vol. 21, Oxford Internet Institute.

[5] Chandaka B. , Mandapati V. and Vedula V. , 2018, Efficient Association Rule Mining for Retrieving Frequent Itemsets in Big Data Sets " , CJAST.39546 , PP.1-14.

[6] Chitra1 and Maheswari, 2017, A Comparative Study of Various Clustering Algorithms in Data Mining, K. Chitra et al, International Journal of Computer Science and Mobile Computing, Vol.6 Issue.8.

[7] Galathiya, Ganatra, and Bhensdadia, 2012, Classification with an improved Decision Tree Algorithm, International Journal of Computer Applications (0975 – 8887) Volume 46– No.23.

[8] Gunasundari and Karthikeyan, 2012, A Study of Content Extraction from Web Pages Based on Links, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.3.

[9] Himani Sharma1 and Sunil Kumar, 2015, A Survey on Decision Tree Algorithms of Classification in Data Mining, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

[10] Jharna M. , Sneha N. and Shilpa A., 2017, Analysis of agriculture data using data mining techniques: application of big data, J Big Data DOI 10.1186/s40537-017-0077-4.
[11] Lourdu C., Jayanthy, and Sakthivel, 2016, Implementation of Different Techniques of Web Data Mining through Cloud Computing Technologies, Volume 6, Issue 6, June 2016 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
[12] Manisha R., Mohod, and Thakare, 2015, Various Data-Mining Techniques for Big Data, International Journal of Computer Applications.
[13] Neha G. and Saba H., 2011, A Heuristic Approach for Web Content Extraction, International Journal of Computer Applications (0975 – 8887) Volume 15– No.5.
[14] Nelofar R., 2017, Data Mining Techniques Methods Algorithms and Tools, Vol.6 Issue.7, International Journal of Computer Science and Mobile Computing.
[15] Pranit B. and Sheetal D.,(2018), Web Data Mining Techniques and Implementation for Handling Big Data', IJCSMC, Vol. 4, Issue. 4, April 2015, pg.330 – 334.
[16] Rajkumar D. and Usha S., 2016, A Survey on Big Data Mining Platforms, Algorithms and Handling Techniques, International for research in Emerging Science And Technology, VOLUME-3, SPECIAL ISSUE-1, NCRTCT"16.