Knowledge-Based Analysis of Web Data Extraction

Marwah N. Abdullah
University of Kirkuk
marwah.nihad@uokirkuk.edu.iq

Alaa Hassan
University of Kirkuk
eng.alaa.hassan@ieee.org

Nadia Naef
University of Kirkuk
nadia.ibra@uokirkuk.edu.iq

ABSTRACT
In this research, the field of mining has organized the content across the Web by providing the models and techniques of working to achieve the integration of knowledge in a mechanism so that these models are designed to represent human knowledge in the form of structured language through the concepts of modeling tools. Various webs used to obtain data from different sites may seem a little complicated at first, where we studied in this research the exploration of data on the Web. The data is analyzed and the following extract used the Web information extraction technology. They are extracting the information from pages through using a program designed in the Java language, this has been implemented by checking every page of your website, then added the extracted information to their database. Documentation Web has many different formulas formats, such as HTML pages and other formats. Data web is an extracted function to detect the state of the web pages contents, if they are hacker pages or not, where evidence is imported to CSV. Next test data uses web content software depending on the decision tree mining algorithms and is implemented in Weka.

Index Terms— Applying data mining algorithms, Classification, Website dataset, Web data extraction.

I. INTRODUCTION
Data mining includes the use of data analysis implements to determine previously uncertain, and associations in huge information. The ease of making information technology and the general use of it obtainable to the big size of information make the concern of great data on the internet have been debated, in terms of the existence of these arbitrary sources and feasibility. Once we talk about the vast information, we are talking about the extents that cannot be abstracted from various types and sources of data. Information is everywhere and this leads to increase the required evolving tools by considering the potency to examine the data and mine information and knowledge from them. Traditional statistical patterns cannot deal with this huge amount and they use intelligent tools to report this information. The objective of data mining system is to conclude knowledge from the great amounts of records by relying on calculated algorithms, which are built on data mining and the result is from many sciences, such as math, logic, statistics, science learning, artificial intelligence and expert systems, knowledge of the machine, and pattern recognition knowledge. The
significance of Web Data Extraction methods relies on the detail that a huge of information is always shared. Web Data Extraction schemes permit to gather this information with restricted human strength and that is one of the smart and non-traditional sciences. Many methods of mining data from the Web have been planned to solve certain problems. Data mining finding big data includes the procedure of complex arithmetical algorithms. Web extracting might be a basis for data mining. Data Mining can gather every basis of data and if that process needs data obtainable from the web then web extracting might be one of the approaches to obtain such data. The Data mining is one of the positive results to investigate enormous quantities of information, and now turning it than objective gathered and unintelligible information into valuable info that can be hard-done and operated later. I have seen data mining various interests in the step of trying to develop a scalable and adapt to the increasing amounts of data algorithms in the search for meaningful cognitive patterns. A package has grown from algorithms and software and dramatically to the extent that the expansion has made it difficult for workers in this field, and trace technologies available to solve a specific task. So that data mining aims to extract knowledge of the content and structure of the web. The method of mining information from Web pages is also named as Web Data Mining. It uses data mining methods in order to achieve interesting and possibly beneficial information from web data. World Wide Web or the biggest dataset comprises different data that we would like to employ for our needs. Relying on data mining algorithms and web forms data and knowledge discovery from the reality of all of the data available on the Web.

II. WEB CONTENT MINING

This pattern of types of exploration on the web to discover the knowledge of the reality of the content of pages and Web sites analysis works, involves exploration and analysis in this style on the sites and content classification according to their themes, as well as involving this category knowledge discovery from the reality of the analysis of the comments and feedback generated by readers and beneficiaries out indicators of interest can invest in various aspects of knowledge, it should be noted that this does not apply to data mining as it is not available in the database tables ability add comments or feedback to the content. Information Extraction has the aim of transforming a set of web pages [1]. The operation of analyzing and collecting information for web is precisely associated to data mining. Relational data mining and extraction models used to determine associations and forms in text.

III. DATA MINING ALGORITHMS

There are various approaches to methodology this problem: association, classification, clustering, etc.

1. Associations: It means the discovery of the correlation relationships among a group of elements. Association Rules are resulting from frequent itemsets [2]. Association support and assurance methods gave good results in several fields. Association rule mining has a widespread series of applicability such Website navigation study. It is easy to implement and practice.

2. Classification: It is the greatest usually practical data mining technique [3]. The classification process is proficient of handling a
wider variation of data and is developing in popularity. The classification includes of predicting an assured outcome depended on an assumed input. So as to calculate the result, the system practices a training set holding a fixed of characteristics and the distinct conclusion. Classification analyzes a usual of training information and builds a sample for all categories depending on the structures in the data, such as Decision trees and lists, classifiers of instance-based, support vector machines, logistic regression, multi-layer perceptron’s algorithm, and Bayes nets. Between different kinds of knowledge depiction current in the classification usually uses calculation rules to precise knowledge.

3. Clustering: It is the definition of built-in data clusters, where the assembly is a set of similar data with each other. The goal is that the objects in a group will be similar [4].

- Decision Tree Algorithm:

A decision tree is a usually used of the data classification algorithms. It is graphic illustration described by tree diagram [5]. Decision trees algorithm are leaves that organize features through categorizing them depending on the feature standards. Every node in this algorithm characterizes attribute in an instance to be categorized. All branches epitomize a value that the node can accept. Attributes are ordered initial at the source node (root) and arranged depended on their attributes. Decision trees scheme can simply be changed into IF-THEN instructions [6]. This is algorithm utilized to classify the features. Wherever the input to the algorithm usual of data and the output is categorized over which you can organize of new data not previously applied as a sort of calculation for this new data. This classifier in the method of tree construction thus called decision tree can be improved into arranged of rules thus called decision rules. This is algorithm used Divide and Conquer principle the idea of splitting the problem into pieces and solves all of them separately and then is assembled solution. It was established decision tree based on the best choice attribute property can be divided training set so that at least the depth of the tree at the same time as the classification of the data correctly. Finally, decision tree constructed is efficient [7]. Decision Tree Comparatively fast matched to other classification models. It is easy and simple to know classification procedures.

- Exploration Web

The World Wide Web, or simply the Web, represents one of the largest sources of information in the world. We think that any subject has become probably exists on a page on the web. Information on the web comes with different shapes and types, such as text documents, images, and video clips. However, extraction of useful information, without the help of some web tools, is not an easy process. Here comes the role of exploration Web, which provides tools that help us to extract useful knowledge from web data. In this paper, we have used these techniques for the detection of secure pages. Perfect information is the major requests for perfect data investigation. Supposing well data is accessible, the following phase is to select the most suitable procedure to exam the data. The data predictor would match or association available methods in order to find the greatest possible outcomes.
- **WEKA Data Mining System**

WEKA is Waika to Environment for Knowledge Analysis. It’s a data mining machine learning mechanism advanced via, New Zealand, Computer Science Department, University of Waikato. It is free data mining software available [8]. Weka is a group of machine learning systems for data mining jobs. The processes can both be used directly to a dataset or applied from your private Java code. Weka includes tools for information pre-handling, regression, clustering, classification, visualization, and association rules.

**IV. RELATED WORKS**

- **Research Problem**

The problem for research in the detection of non-hacker-protected web pages and the existence of a huge amount of data in terms of scientific and technological methods used in dealing with this data and evaluated and make decisions on the data obtained which determines safe web pages or unsafe web pages.

- **Basic Models**

**Dataset:** A fixed of data items; the dataset is a precise simple idea of machine learning. A dataset is approximately correspondent to database table or a two-dimensional worksheet. In WEKA, it is executed by the Instances.

**Instance:** consists of a number of attributes.

**Item set:** A set of items that regularly show collected in a coefficients data set.

**Attributes:**

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

**Confidence:** The confidence is well-defined by a qualified possibility Sureness(X=>Y) = Provision (XUY)/ Provision (X) =P(Y/X).

**Lift:** It is the relation of the possibility that L and R happen collected to the several of two distinct possibilities for L and R. This paper illustrates the used of java to examine HTML pages and to get page content, then extracted the web pages content into "webhacker.arff" file. The example data set applied for this example is depended on the "webhacker.arff" by applying a decision tree algorithm with WEKA program. This paper assumes that appropriate data preprocessing has been performed. The URLs and the data extracted from these URLs are shown below in Figure 1 and Figure 2 as a tabular form:

**Figure1:** Analysis URL (HTML pages) to retrieve page content using java
The implementation includes:

- The data set name – “webhacker.arff”
- Number of cases in the relation (row) = 191
- Number of characteristics (column) in the relation = 3

Attribute description:

Id: identification number
Words: string
Hackers :{ yes, no }

Load data set in Weka:

- Load the bank data in Weka (click on Preprocess and then on Open file... "webhacker.arff").

- Decision tree algorithm to the dataset "webhacker.arff"
  - Once studying the data applying the provided algorithms.
  - Investigate the data with C4.5 algorithm relating J48, WEKA’s employment of decision tree expert. The example data utilized in this workout is the web information from the file “webhacker.arff”.
    - After you have your data set occupied, totally the tabs are accessible to you. Click on the ‘Classify’ label
    - Click on ‘Choose’ key in the ‘Classifier’ container just lower the tabs and choice C4.5
    - In this proposal, you will estimate classifier depended on how fine it expects
    - When training set is complete as shown in Figure 3 and Figure 4:
The ‘Classifier’ output area describes the results of training and testing as following:

The fixed of amounts is resulting from the exercise data. In this situation, only 88% of 190 training instances have been categorized properly. This shows that the outcomes found from training data are not confident matched with what might be acquired from the free test set from the equivalent basis. Furthermore to ordering inaccuracy, the assessment output measurements resulting from the session possibilities allocated via the tree. More precisely, it outcomes means outcome fault (0.1) of the possibility approximations, the root mean squared fault (0.3) is the square root of the quadratic loss. The mean absolute error computed in an equivalent manner through applying the absolute in its place of squared modification. The cause that the faults are not 0 or 1 is because not completely training instances are classified properly.

After that applying the clustering algorithm and gets the following results:

The visualize assignment clustering is as below:
IV. CONCLUSION

In this research, website data has been analyzed, extracted, evaluated and classified into web pages which do not represent hackers and dangerous Web pages which represent hacker and these rely on data derived from these pages. After obtaining the information, this data was analyzed using valuation Data mining techniques to obtain accurate results in the classification by using decision tree and clustering algorithms. It has found that the algorithms used are better than the others in performance. In the future, we develop the program to extract all the links from a given web page, so as to extract more data and get the best results in the exploration of these data.

REFERENCES


SMC Hiroshima Chapter, Hiroshima University, Japan, PP.208-212.

