International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 11 • Issue 1 pp. 602-606 Jan 2019-June 2019 www.csjournals.com

Auto Approach for Extracting Relevant Data Using Machine Learning

[1]Dipali Shete, [2]Dr.Sachin Bojewar

[1]PG Scholar, [2]Professor & DAO, Department of Information Technology VIT, Mumbai

Abstract: The current debate revolves around how to obtain relevant data from the Internet. In data mining web information extraction is an important area which examines demanded data. The web document contains relevant data in a structured or unstructured format. It means that abstraction of required data from web document is in HTML format. This extracted data may be used for retailer business and for data analysis purposes. For Online retailer websites, product data is critically important. The manual extraction of such data from multiple manufacturers' websites is complicated as well as time-consuming. To resolve this difficulty we are going to develop such systems that will automatically extract retailer required data (Relevant Data) of product descriptions over various sites and determining the patterns or features present in that information based on user-provided information. We propose a system that will automatically extract the user required data and deliver it to the user through PKV structured in .txt format. We define a fully visual and interactive user interface with a new approach and technique using scrapping tools and machine learning algorithms.

I. INTRODUCTION

With the growth of the internet, Today we can acquire an increasing number or quantity of information quickly, how to systematically obtain relevant information from the web page is challenging in web data mining. The information present on the internet is very big and it is not possible to collect relevant data from different websites manually in the information extraction process. The web data is in structure and unstructured format which means with relevant information on web page it contains noise or unwanted data such as advertising images, videos, copyrights, etc. Which
make it difficult to find particular and valuable data from a web page for the web users many data mining methods are available today to extract information from the web page. Data mining means collecting information from a huge amount of data from anywhere.

Data mining is extensively used in various areas. There is a number of commercial data mining systems accessible today and yet there are many difficulties in this field. Data Mining has a large-scale application in Retail Business because it receives an extensive amount of information from customer purchasing, history, sales, goods, sellouts, consumption, shipping, and services. It is essential that the quantity of data gathered will continue to increase rapidly because of the growing ease, availability, and demand of the web. Data mining in the retail business supports in knowing consumer buying patterns and trends that guide to recovered group of consumer service and good consumer remembrance and satisfaction. So extracting relevant information from web page is an essential topic and it has attracted much attention. According to statistics, existing web data extraction methods can be reviewed as follows wrapper based method, Dom tree rule-based method and manually obtaining information.

We developed the system which will be obtained relevant product data from a particular manufacturer website based on template matching. Information on the web is saved in the form of the templates. Extraction in the declared system is concerned about data that are saved in template format. The Paper gives a whole description of all the ideas and techniques of auto relevant Information Extraction from Web Pages. The declared system will automatically obtain data from web pages with varying content and templates produced by the user. The web crawler is a script that crosses the website web pages based upon web indexing which will be downloading the pages and following the links from page to page. Then determined data based on user-provided information from the downloaded pages using Jsoup. After the collection of data from different websites next step is data preprocessing, cleaning, filtering and streaming data. Jsoup is nothing but java libraries that used to deal with real-world HTML. It consists of a very useful API for creating and collecting data, by utilizing the best of CSS, Jquery and DOM-like methods. Jsoup implements the WHATWG HTML5 term further parses HTML to the same DOM. This allows capacities like parse HTML from URL and finds proper data based on DOM Traversal or CSS selector, then we are

using the Naive Bayes Classifier which will examine data based on user-provided information of required data. Using this proposal automatically obtains user required data and presents it to the user through PKV structured .txt, DAT format file.

II. LITERATURE SURVEY

This section gives an overview of the research carried out so far in this area and the techniques employed in developing an Extracting web data system of the following:

- Extracting text information from web page based on block and Tag-functions.
- Content Information Extraction of Theme Web Pages based on Tag Information.
- Comparative Analysis of Various Methodology to Detect Paragraph from Web Document
- Unsupervised approach for semi-structured data record extraction from multiple pages using tag tree similarities.

1) **Extracting text information from web page based on block and Tag-functions:** In this paper proposes text information obtained based on visual information and tag function. In their method, they first separate a web page into various blocks based on visual information and then extract text information in this block based on the tag function characteristics in the DOM tree and used VIPS algorithm which carries noise in the block and this noise block deletes using visual information and location information of block in page [1].

2) **Content Information Extraction of Theme Web Pages based on Tag Information:** In order to obtain the con- tent information of Theme Web Pages more precisely, this paper proposes a self-learning system based on the tag information by determining the information quantity of different tag indicators. This design predefines many tag information indexes and coefficients index to calculate a type of tag information number of the web pages in turn, and then the candidate content of Web pages is in the tag with the maximum data quantity. To improve the versatility of the system, they add the adaptive and flexible coefficient weight in the calculation formulas of tag information quantity. With the increasing of data be provided, tag collections, index value and the information quantity results are combined into the learning database to adjust the weight of the coefficient factor [2].

3) **Comparative Analysis of Various Methodology to Detect Paragraph from Web Document:** We can analyze various methodology to detect paragraph from web document for that tenacity we see the information about web document and obtain the web contents in the new web page so, in this paper used HTML parser, and HTML web Page and Eclipse. They take partial contents of web pages. They introduce an information aggregation scheme that obtains partial contents of web pages. They derive paragraphs from web sites and perform work on them. For that purpose, the system uses the parsing procedure in which parse HTML document, links of URL file. Java queries to use for obtaining and managing data. They introduced a system for aggregating partial data across multiple web pages and also used the jsoup library for paragraph extraction from an offline HTML Document. The suggested method retrieves web pages based on a users query, segments those pages into partial areas for each user. They performed the introduced method as a prototype system [3].

4) **Unsupervised approach for semi-structured data record extraction from multiple pages using tag tree similarities:** In this paper, they present a novel unsupervised procedure for data records extraction from various similar web pages using tag tree similarities. Obtaining the data records from multiple web pages consist of the following orders. They first recognize the related web pages from the web source. Next, they create the DOM tree for related web pages using HTML parser. Then examine two or extra web pages to reduce unwanted areas such as header, menu bar, and navigation bar, advertisements, etc and find the region carrying data records also indicated to as data region. Then traverse sub trees of data region to obtain specific data documents and store them in the expected form such as XML. The main giving of this paper is in generating a fully unsupervised algorithm for obtaining both structured and semi-structured data records from various related web pages. The proposed system can obtain important data records from several commercial web sources more accurately. Therefore, it can assist as a tool for combining data from different commercial websites [4].

III. PROPOSED SYSTEM

The objective of proposed system is auto approach for extracting relevant data using machine learning is to auto extracting and displaying the relevant data based on input given by the user by processing it through various mediums. As shown in Figure 1. Following components are implemented in this system.

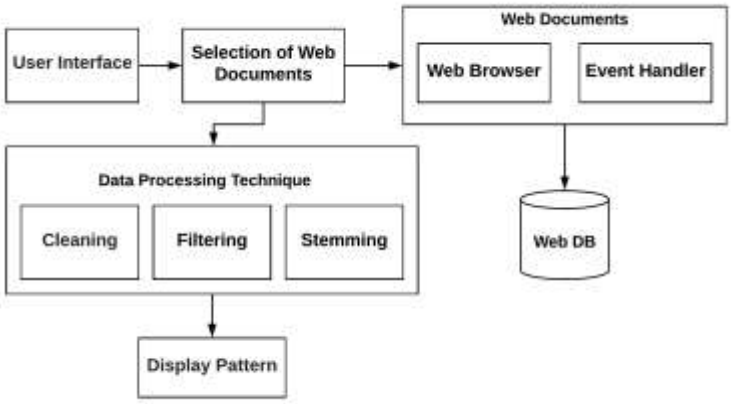

Figure 1: Proposed System

1) **User interface:** User interface is used to take the inputs from user in text format i.e. short descriptions, long descriptions, Xpath, Tag Class Name, I'd or complete data.
2) **Selection of Web:** In this phase, based on input values given by the user, required data extracted in raw format using the Naive Bayes method which is a simple Bayesian classifier is based on Bayes' theorem from probability theory. This raw data from web document is further given to Data pre-processing phase to get the relevant data.

3) **Data Extracting and Pre-Processing:** Data Extraction is the process or act of capturing web data that is either semi-structured or unstructured from web pages and turning the raw data into analyzable or table style formats. In our project we are extracting data from various websites or WebPages.

In Text Pre-processing, we need to extract relevant information from Web sites or Web pages so here we are using Crawler and parser for extracting the information regarding blogging sites, crawled and parsed for information collection and for text processing on those data. **Data Pre-processing:**

After collection of data from different websites next step is data pre-processing. Data Preprocessing is necessary to transform raw data into understandable format. In this we are having three steps:

- Data Cleaning: Data cleaning in data mining is the process of identifying and eliminating corrupt or inaccurate data from raw data from web documents.
- Data Formatting: Formatting the data to make it suitable or in structured format to process it further.
- Data Stemming : Stemming is a process where words are reduced to a root by removing inflection through dropping unnecessary characters cleaning, filtering and steaming data done step by step and process followed are:

For extraction of the relevant data from the irrelevant one, we are using Jsoup. To deal with real world HTML there is a java library named Jsoup. It gives a very comfort API for manipulating and extracting data, by using the best of DOM, CSS, and jQuery like methods. Jsoup implements the WHATWG HTML5 specification and parses HTML to the same DOM as modern browsers do. The functions of the Jsoup are:

- Scrape and parse HTML from a URL, file, or string.
- Find and extract data, using DOM traversal or CSS selectors.
- Manipulate the HTML elements, attributes, and text.
- Clean user-submitted content against a safe white-list, to prevent XSS attacks.
- Output tidy HTML [6]

IV. ALGORITHM USED

Naive Bayesian Algorithm: An algorithm which normally used in machine learning for data classification. Naive Bayesian is set of classification algorithm based on Bayes theorem. In which some group of algorithms shares the same principle, in which every feature is independent on the value of any other characteristics. An example if we consider Orange as fruit. It must have characteristics like orange color, round in shape and 3-4" diameter. Naive Bayes algorithm examines each these characteristics to produce independent to probability which specify fruit is an orange despite any relationship between characteristics. Sometimes characteristics not always independent at this time Naive Bayes fails to obtain proper results which is a weakness of this algorithm that's why this algorithm is known as "naive". In our proposed System web documents are obtained using this algorithm. Web documents are extracted based on the steam formation of a word which is produced in the training phase by a user. The basic steps in the naive Bayes method are as follows:

– Training phase:

- The individual stem words occurring in all the training documents in the training set are identified.
- For every document in the training document set form the characteristic vector and also save the same in the knowledge base along with suitable indexes.
- For each index, probability gets calculate.

– Testing phase:

- Identify the individual stem words arising in a given text document.
- For this document form the characteristic vector. Calculate the probability for this document given each index.
- For this document, for every index in the set of indexes, calculate the probability and normalize the same with Bayes theorem, this value is the weight of this index.
- Select the indexes with a weight higher than a predefined inception as the candidate indexes for this document [5].

V. METHODOLOGY

Jsoup and Web crawling:

Jsoup is a Java library that is used for working with real-world HTML. It gives a very beneficial API for obtaining and handling data, utilizing the best of DOM, JQuery, CSS, like methods. Or, in different words, it is a Java library that provides you to scrape or retrieve specified information from websites using jQuery, CSS methods. For finding new or updated pages for indexing we used web crawler is a Script that travels the Web. The Crawler starts from roots websites or a large range of public URLs (also recognized as the frontier) and searches in depth plus width for hyperlinks to obtain. Web crawling in easy terms is obtaining data from the websites. You need such information to examine and obtain meaningful and precise results. The web is overfilled with a variety of data and how we use it to optimize our business decision is part of a Data Scientist's work. We are viewing at a Java API, Jsoup which will be utilized to obtain data from websites [6].

Data Mining:

Data mining is a process of identifying Information from a data warehouse. This information can be classified in different rules and models that can help the user to study collected data and predicted decision processes. The main database of any organization is known as Data warehouse, wherever all data is stored in a particularly large database. Data mining is a process that is used by the user to get valuable knowledge or information from raw data. Software's are performed to look for required patterns in a huge quantity of data that can support the business to discover about their customers, predict behavior and develop marketing strategies. Web mining is truly a field of data mining associated with the information obtainable on the internet. Users use various search engines to fetch their expected data from the internet, that information and user required data is determined by mining method called Web Mining. Web mining is rapidly growing very powerful due to the size of text documents growing over the internet and taking important patterns, knowledge and informative data is very difficult and time-consuming if it is done manually. Structure, Usage, content is included in information found by Web mining. Some processes included in web mining are Information Retrieval and information extraction. Ma- chine Learning is support means that helps in obtaining data from the web. Machine learning can enhance web search by understanding user behavior. Several machine learning methods are applied in a search engine to produce brilliant web service. It is much more productive than the traditional way i.e. information retrieval. It is a process that holds the capacity to study user action and improve the performance on a particular task.

DOM Tree:

The standard for locating, extracting HTML and XML components from the web page is Document Object Model (DOM).The Document Object Model is a stage and language-independent approach for representing purposes in HTML and XML documents. The Document Object Model (DOM) is an application programming interface (API) for actual HTML and XML documents and it is a parser that parses the HTML and XML document in series to form a tree-like structure. Objects in the DOM tree can be managed by using functions, rules on the objects which are present on the web page. DOM tree allows a dynamic path of programs and scripts to update the content, style and, building of the page. DOM secures the rational structure of documents and the way, a document is obtained and managed. With the help of Document Object Model, a tree structure can be formed for the complete web page including relevant and noisy node. With the Document Object Model, programmers can create documents, navigate their formation, and alter, add, delete components and content in the tree-like structure.

HTML parser:

HTML Parser is a Java library utilized to parse HTML in each a nested or linear fashion. HTML Parser produces its integrity in design, speed, and capacity to manage running real-world HTML. The two fundamental use-cases that managed by the parser are transformation and extraction. While earlier versions focused on data extraction from web Documents, whereas Version 1.4 of the HTMLParser holds valuable improvements in the area of transforming web documents, with simplified tag editing and creation, and verbatim to HTML() method as output. In worldwide, to use the HTML Parser you are must be familiar with Java programming language because you will need to write a script or code in java language. Although unusual example programs are presented that may be useful as they persist, it's more than likely we will need to develop our own script/Program or change the ones produced script to meet our proposed system. To handle the library, we can add and set either the htmllexer.jar or htmlparser.jar to your classpath during compiling and running. Extraction contains all the information retrieval applications that are not expected to save the source page. These covers use like

- Text extraction, for use as input for text search engine databases for example.
- Link extraction, for crawling by web pages
- Screen scraping, for programmatic information input from web pages.
- Resource extraction, gathering images or sound.
- A browser front end, the introductory stage of page display.
- Link checking, assuring links are valid.
- Site monitoring, monitoring for page variations be- yond simple differences [3][6].

VI. CONCLUSION

Thus, in the proposed approach, I am going to implement a system that can be used to crawl the web documents from various manufacture websites and store them in PKV structured in .txt format fie. This can be done using the Naive Bayesian algorithm. The crawling framework can work well using Jsoup, HTML parser and DOM tree. Obtain data based on user-provided data this data can be actual product information, Xpath, Tags. Such obtained relevant data applied for web indexing, data mining, data harvesting, monitoring modifications to websites and differences to content information. Online retailers have never seen back in spending millions of dollars and comparable efforts for establishing their brand appearance online and earning client commitment; the clients who represent the lifeblood of the retail business. Still, there are retailers who work really difficult to collect data from the web that can help them steal clients from the competition and permanently win their businesses.

VII. REFERENCES

1. Dingrong Yuan, Xiaohu Yang, Huiwen Fu, & Xue Nong. (2012). "A new algorithm: extracting text information from webpage based on block and tag-function". IET International Conference on Information Science and Control Engineering 2012 (ICISCE 2012).doi:10.1049/cp.2012.2437.
2. Wang, J., Wu, J., Zhang, Y., & He, G. (2014)."Content Information Extraction of Theme Web Pages Based on Tag Information". 2014 Seventh International Symposium on Computational Intelligence and Design. doi:10.1109/iscid.2014.257.
3. Narendra. M. Jathe, Ku.Nayana B. Neware, Hemant Ma- halle (2017). "Comparative Analysis of Various Methodologies to Detect Paragraph from Web Document". 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939.
4. ALEEM ANSARI, HEMALATA VASISTHA (2015) "UNSUPERVISED APPROACH FOR SEMI-STRUCTURED DATA RECORD EXTRACTION FROM MULTIPLE PAGES USING TAG TREE SIMILARITIES" Proceedings of 34th IRF International Conference, 9th Aug 2015, Pune, India, ISBN: 978-93-85465-73-4.
5. Yong Wang, Hodges, J., & Bo Tang. (n.d.). "Clas- sification of Web documents using a naive Bayes method. Proceedings". 15th IEEE International Conference on Tools with Artificial Intelligence. doi:10.1109/tai.2003.1250241.
6. https://jsoup.org/ - JSOUP/HTML parser