International Journal of Emerging Trend in Engineering and Basic Sciences (IJEEBS)
*ISSN (Online)*2349-6967
Volume 2 , Issue 1(Jan-Feb 2015), PP424-429

History Visualization using Information Technology as a Tool

KiranChaudhari[1], VarshaThorat[2], Nikhil Mandlik[3], HarshadKedia[4]

[1 2 3 4]*(Dept. Information Technology, Dr. D. Y. Patil College of Engineering)*

Abstract :- This paper presents an automatic approach to intelligent data mining of unstructured text and innovative visualizations after conversion to structured text. Unstructured text is obtained from URL's and converted to structured text using Natural Language Processing. Then the rules engine identifies signal from the noise. Post that curation engine cures the data which is then sent to Visualization engine. One of the major goals is to provide Visual education. The example chosen for the same is Visual History of the Maratha Empire. This approach can be extended easily for other empires and other large scale events. Experimental results show that above approach can achieve the high detection accuracy, lower detection time and performance with a small sample of the classification model training set.

Keywords :-Data mining using web scraping, Natural processing language(NLP), Data Visualization

I. INTRODUCTION

A vast amount of historical data is present on various data sources e.g. Wikipedia, Military History, Newyork Times etc. Most of this data is in plain text form. However, history did not happen on pages. History happened on places / countries on maps. Thus there is huge scope for improving the way history is learnt by students. Data mining and visualization is fundamental in automating and visualizing historical information from multiple data sources using Google maps and Google charts. The system will automate aggregation of historical data from multiple data sources and then visualizing it on Google maps and other Google charts e.g. scatter chart, bubble chart, pie chart, line chart etc. All the visualizations will be deployed on Google cloud there by enabling the world to visualize history.In the above paragraph, total process of the planned system is given below in short. This system is using intelligent data mining through web scraping, later conversion of unstructured text to structured text and innovative visualizations. There are many sources available on the internet that gives information about everything. Insome study scenarios today's generation knows that what happened in history, but do not know where the corresponding locations of history that are shown on Google Maps,which are widely usednow. So, this system targets these above mentioned areas, and uses technologies like Data mining, Webscraping, Natural language processing, Rule engine, Curation engine, Visualization engine,,etc.. The first example chosen is to visualize the history of the Maratha Empire. It will be shown on the maps with current maps provided by Google.Along with this the events happened on those places will be shown as well.

II. LITERATURE SURVEY

The system's concept is based on the IEEE paper. The authors have developeda new system for Cancer Registry and Regional Cancer Network Integration. It is fundamentalto improve validity and timeliness of data diffusion when both the number of sourceslinked and the number of variables registered are on the rise Aims. It contributes to shorteningall phases of cancer registration, and including linkage with the external sources, coding, qualitycontrols, data management and publication and analysis of results. Integration in the oncologynetwork and secure Web access allowed us to design with clinicians innovative populationbasedcollaborative studies. Geographic analysis system enabling to develop sophisticateddynamic geo-statistic tools[1].

2.1. Data Mining using Web Scraping

Web scraping is the latest computer software technique for web data extraction of information from URLs and central websites provided by the admin. This kind ofprograms does provide human exploration for the World Wide Web by either implementing low-level Hypertext Transfer Protocol(HTTP), or embedding the complete central sites provide by the user such as for Google.The technique ofWeb scraping relates to web data extraction which indexes information on the web usinga bot or web crawler and is a universal technique adopted by most search engines.Web scraping focuses more on the transformation of data to be unstructured on the central site, which is inthe HTML format, into data which is of structured type that can be stored and analyzed in a central local databaseor spreadsheet. Web scraping is also related to web configuration, which decides how human browses with the help of computer software. Uses of web scraping include online price comparison, contactsscraping and in many other fields the technique of web scraping is used[2].

III. ARCHITECTURE

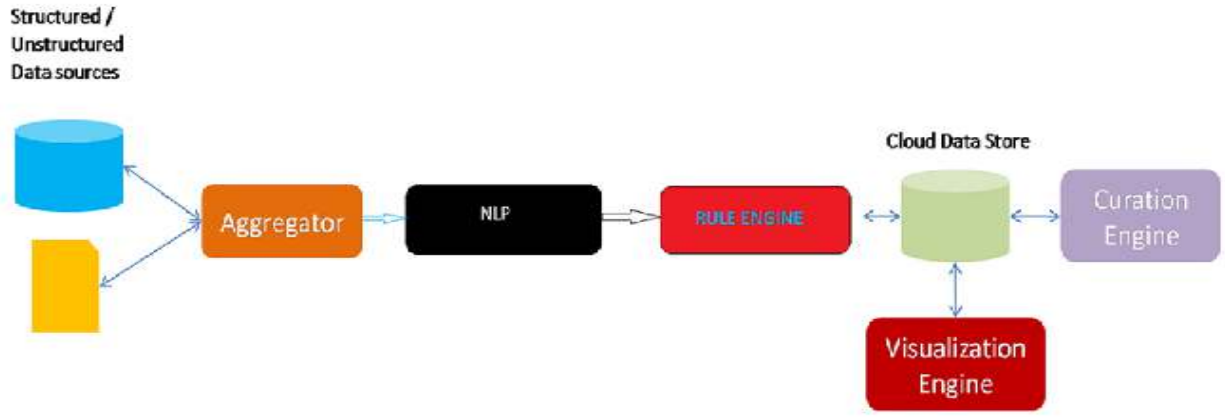

Fig.1. System Architecture

3.1. Data Mining using Web Scraping

First of all the data mining will be done to obtain related structured or unstructured data. Web scraping is the process of automatically collecting information from the World Wide Web. Its a field with active developments sharing a common goal with the semantic web vision, an ambitious initiative that still requires breakthroughs in text processing, semantic understanding, artificial intelligence and human-computer interactions. Web scraping, instead, favourspractical solutions based on existing technologies that are often entirely ad hoc.Web scraping is closely related to web indexing, which indexes information on the web using aboot or web crawler and is a universal technique adopted by most search engines. In contrast, web scraping focuses more on the transformation of unstructured data on the web, typically inHTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web scraping is also related to web configuration which decides how the human browsing is done with the help of computer software provide to them.

3.2. NLP: Natural Language Processing

After the unstructured data obtained, it should be processed to obtain clear language. Andalso if it extracted from the Wikipedia then from the page there is too much data that is notneeded. So to do so NLP is used.Natural language processing (NLP) is a fieldwhich deals with the interactions between computers and human (natural) languages.There are many challenges involves in natural language that is, enabling computers to derive meaning fromhuman or natural language input, and others involve natural language generation. After thewhole process is done, the data is refined and can be shown on required place [3].The Natural Language Processing unit works on the data gathered from sources like Wikipedia.For example when we will collect data about one of the battle and we will find and locate it onto the Wikipedia then with help of Rule engine, on the particularly collected data NLP willsimplify the data and convert it according to its aspects of rules and regulations [4].

3.3. Rule Engine

While collecting the data from the external sources like Wikipedia, on the particular webpage there will be different types of data. So the system needs only some part of that information.Before giving the data to the NLP unit Rule engine retrieves only needed blocks while excludesunwanted data. It will collect the text data from Wikipedia web page and will exclude other likeother text matter of links, ads.

3.4Curation Engine

This system will always collect data from external resources. But in some scenarios the data might not be available. So the Curation Engine will provide the facility only to the admin to edit or create and upload new data for some events. Unlike the Wikipedia anybody will not able to update the data. With the help of forms admin will upload data.

3.5Visualization Engine

Data visualization or data visualisation is a modern branch of descriptive statistics. It involvesthe creation and study of the visual representation of data, meaning ”information thathas been abstracted in some or the form, for the units ofinformation”. The main advantage of data visualization is to communicate information clearly andefficiently to users via the information graphics selected, such as tables and charts[5].

IV. ALGORITHM FOR WEB SCRAPING TO EXTRACT DATA FROM HTML

```
function webscrape() {
   contents = read_content(url)
   parse_html(contents)
}

function parse_html(contents) {
    tokens = tokenize(contents)
}

function tokenize(contents) {
    do {
       node.starttag = identify_start_tag()
       node.attributes = identify_attributes()
       node.text = identify_text()
       node.endtag = idenify_end_tag()
```

```
        node.parent = identify_parent()
        insert_into_tree(node)
    } while(node.endtag != '</html>')

}
```

4.1. HTML parsing

Parsing can be separated into two sub processes - lexical analysis and syntax analysis. Lexical analysis is the process of breaking the input into tokens. Tokens are the language vocabulary - the collection of valid building blocks. In human language it will consist of all the words that appear in the dictionary for that language.Syntax analysis is the applying of the language syntax rules.

4.2. Tokenizing

The parser will usually ask the lexer for a new token and try to match the token with one of the syntax rules. If a rule is matched, a node corresponding to the token will be added to the parse tree and the parser will ask for another token. If no rule matches, the parser will store the token internally, and keep asking for tokens until a rule matching all the internally stored tokens is found. If no rule is found then the parser will raise an exception. This means the document was not valid and contained syntax errors.

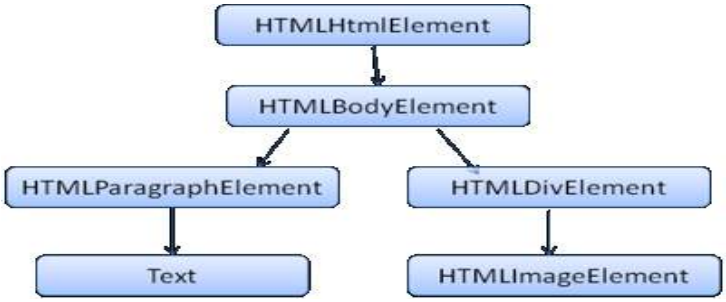

Fig.2. Tokenizing

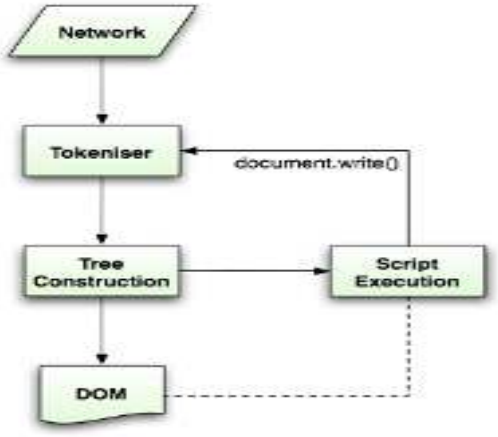

Fig.3. Tokenizing

- NLTK grammar type for each word extraction e.g. nouns, verbs etc
- Anaphora resolution to convert pronouns into nouns
- Simple matching to identify key phrases and relationships e.g. Inheritance is part of oops. Thus oops is the root node and inheritance comes under that node.

V. MATHEMATICAL MODEL

- Span of the empire(S)

S=D1-D2

- Number of battles, Number of kings

Sum of battles and sum of number of kings.

- Battle success rate(R), Battle failure rate (F)

  R=W / N

- Growth rate(G)

  G= (CA - PA) / (PA)

- Max land area and % of world area

  % world area =A * 100 / WA

- Performance(P) score of each king

  P = W1 *N + W2 * G + W3 * R

VI. CONCLUSION

From the above given paper we conclude that visualizing history will be quite easy as history does not happen on pages because when it comes to studying history from books it is not so worth to study and generate interest.As Now-a-Days online Google maps are available on Smartphone and other multimedia device so it becomes quite easy to visualize history because if the history is visualize on Google maps it will efficient and it will interesting also and that's the main advantage of visualizing history on Google maps. Another main goal to visualize is that there will be a central site provided so that the history enthusiasts can view history at any place and time as they wish to do.The other main advantage of this is that the NRI or foreigners who are history enthusiasts can visualize history and they are not cheated when it comes to visualizing history.The main purpose of our project is it is user friendly and finds different applications in the field of medical sciences, educational purpose,tourism, report mechanism etc.In this paper visualization of historical information will simplify history education. The applications of thissystem will be to improve the teaching scheme which is currently associated with only textualmanner and book restricted. So this system will help to better understanding with the pictorialviews which will be more realistic.

VII. ACKNOWLEDGEMENT

Every engineering student looks toward the final year project as an opportunity by which he can implement the skill that he has eventually nurtured in the year by hard work dedication the milestone of completing the project would have been intractable without the help of few people who need to be acknowledge. We owe this moment of satisfactions with a dear sense gratitude to our internal guide Prof. Dhanashree Kulkarni who guided us at every stage. Whose technical support and helpful attitude give us high moral support. We would also like to extend our

sincere thanks to our H.O.D. Prof. Ravi Patki for his guidance and constant encouragement. We are highly obliged to the entire staff of the information technology department and PrincipalDr. S.D.Shirbahadurkar for their kind co-operation and help. We also take this opportunity to thank all our colleagues who baked our interest by giving useful suggestions and also possible help. At last but not least we are thankful to our friend colleagues and all the people directly or indirectly concerned with this project.

REFERENCES

[1] Fortunato Bianconi, Member, IEEE, Valerio Brunori, Paolo Valigi, Member, IEEE, Francesco La Rosa, and Fabrizio Stracci, "Information Technology as Tools for Cancer Registryand Regional Cancer Network Integration", VOL. 42, NO. 6, NOVEMBER 2012

[2]Vasani Krunal A." Content Evocation Using Web Scraping andSemantic Illustration**",**IOSR Journal of Compute Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16,Issue 3, Ver. IX (May-Jun. 2014), PP 54-60

[3] Steffen Koch, Member, IEEE, Markus John, Michael W¨ orner, Andreas M¨ uller, and Thomas Ertl, Member, IEEE"VarifocalReader – In-Depth Visual Analysis of Large Text Documents",Citation information: DOI 10.1109/TVCG.2014.2346677, IEEE Transactions on Visualization and Computer Graphics

[4]Itziar Aldabe, Montse Maritxalar,"Semantic Similarity Measures for the Generation of Science Tests in Basque",JOURNAL OF LATEX CLASS FILES, VOL. 6, NO. 1, JANUARY 2007

[5] Ying Zhu Georgia State University," Introducing Google Chart Tools and Google Maps API in Data Visualization Courses",Computer Graphics and Applications, IEEE (Volume:32 , Issue: 6 **)**

[6] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, Senior Member, IEEE, and Alejandro Jaimes,"Sensing Trending Topics in Twitter'', IEEE (Volume:15 , No: 6 **)**

[7] Faustina Johnson, Santosh Kumar Gupta, Department of Computer Science & EngineeringKrishna Institute of Engineering & Technology,Ghaziabad-201206, India,"Web Content Mining Techniques: A Survey'', International Journal of Computer Applications (0975 – 888) Volume 47– No.11

[8]Raghunath Rajachandrasekar, Zoya Ali, Srinivas Hegde, Vilobh Meshram andNishanth Dandapantula Department of Computer Science and Engineering,The Ohio State University,"Location-Based Query processing: Sensing our Surroundings''

[9] Alberto CavalloMIT Sloan,"Scraped Data and Sticky Prices''

[10] Govind Murari Upadhyay, Kanika Dhingra(Assistant Professor) IITM, Janakpuri, New Delhi, India," Web Content Mining: Its Techniques and Uses'', Volume 3, Issue 11, November 2013