A NOVEL APPROACH FOR CONVERSION OF SEMISTRUCTURED TO STRUCTURED DATA

B.Suchitra
Assistant Professor, Sri Krishna Arts and Science College
Mail id:suchipradeep@gmail.com

Dr. S. Duraisamy
Professor, Chikkanna Government Arts College Kuniamuthur - 8 Tiruppur
mail id: sdsamy.s@gmail.com

Abstract: The flora of the Internet and the status of net objectivity is that transformation can come from everybody. Web mining is a significant research part in today's world. The Web mining is mostly dispersed into three types. First, the web structure mining, second, the web content mining, and the third it focuses on web usage mining. The web usage mining deals with web usage logs which is used to find out patterns from web. The Web structure mining is used to mine information from the structure of hyperlinks .The web Content Mining is used to extract useful information or data from web page. The web content mining is connected but differs from data mining and text mining. In this paper, the research work motive is to learn about web content mining tools, techniques and the examination is concentrated on semi structured data.

Keywords*:* Web Content Mining, Web usage Mining, Web Structure Mining, Structured Data, Semi Structure Data, Multimedia data, Tools of Web Content mining

I.INTRODUCTION

Web Mining is a division of Data Mining technique which is used to find out and extract information from the web documents. Web mining can be divided into following subtasks. They are

- Resource Finding
- Information assortment and preprocessing
- Generalization
- Analysis

The companies, Organizations and persons are eager to collect information through web data mining which can be used to kindle business and to identify market dynamics and new promotions floating on internet.

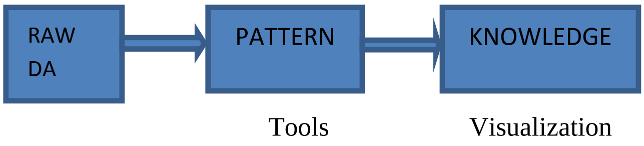

Figure 1.1 : Web Mining Process

The Raw Data can be in the figure of text and images. By using the mining tools, the prototype can be obtained from the raw data and the information can be transformed into knowledge by historical patterns and potential trends. The above Figure 1.1 shows the overview of web mining process.

II. WEB CONTENT MINING

Web Content mining is the mining, drawing out and incorporation of important data from web page content. The content can be a text, image, audio, video, metadata and hyperlink. It also distinguishes personal homepage with other web pages. The Web Content mining is connected but differs from data mining and text mining. The data mining focuses on structured data or structured text. The text mining principally focuses on unstructured data or unstructured text. The Web Content mining is a semi structured data or text and also it concentrates on unstructured data or text. The Natural Language Processing and Information Retrieval are the technologies used in web content mining.

Data / Information Extraction

In this, the structured data are retrieved from web pages which can be from Products and Search results, Value added services Example: Contrast of shopping, Meta search. The techniques used in data/information mining are machine learning. Structured data is simple to extract. Primarily, three approaches are used to extract the data. The primary method is to physically write an extraction program for each website based on observable format patterns of the site. This approach is not scalable for large number of sites. The second method is wrapper induction/wrapper learning. The set of trained pages are manually labeled by user, at first. A Learning system then generates rule for the training pages.

Web Information combination and Schema alike

Here, the different web sites are compared. The different website contains similar information which can be represented differently. This approach focus how to match semantically and also it specifies how to identify similar data. Most of the web pages in the web are likely to be seem as text documents. The researches are closely related to information retrieval and natural language processing. Next researches focused for Web question-Answering

Drawing out Online Opinion Sources

This approach used for customer review of products, forum, and blogs chat rooms. It is used for market intelligence and product benchmarking. This method is used to extract information from multiple sites to provide value added services. Ex: meta search, deep web search etc.

Mining Web to Build Concept Hierarchies/Ontology

This method is getting popular based on ontology. Owl language is used and it focuses for rdf. This method helps to construct the concept hierarchies. The standard method for information organization is concept hierarchy/Category. This is a trendy technique which is used to groups similar search results together in a hierarchical fashion.

Segmenting Web Pages and detect Noise

The web pages mostly contain the content, advertisements, navigation links and copy right notices. This method is used to segment the noise and to remove all the advertisements, links, copy rights and to bring out only the content. The Noisy blocks are removed by using classification and clustering. By using this we will be able to produce much better results.

III.VIEW OF DATA

Data in the web is redundant and many data’s are available in the internet. Data can be viewed and it can be seen in two views. They are information and database view.

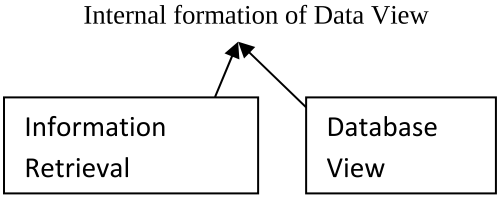

Fig 2: Two Types of View

Information Retrieval

Information retrieval can be center for semi structured documents. The information retrieval uses richer exemplification for feature based on information from the document structure such as html, hyperlinks. The information retrieval uses data mining methods

Database View

It tries to gather the structure of a website or it also transforms a website to become a database. By using database, we can manage the information better and also it can query well in on the web. This can be achieved by finding schema of web documents; building a web warehouse, building a web knowledgebase and also building a virtual database.OEM are mainly used by semi structured data by a labeled graph and the process typically started with the manual selection of websites for content mining.

IV. WEBCONTENT MINING METHODS

Data is everywhere and anytime we can access data over the net. But most of the data in the internet is redundant. The data can be categorized as surface web and deep web. The Enhanced work performed by search engine is web content mining. The two approaches used in web content mining is agent based approach and database approach. The agent approach is focused on three types. They are

(i)Intelligent find Agents:

Based on a particular query, the intelligent search agent automatically searches for information

(ii)Personalized Web Agents:

Based on user preferences it will discover documents related to those user profiles (iii)Information Agents:

It uses number of techniques to filter data according to the predefined instructions

In Database approach, it focuses on complete database which comprise of schemas and attributes with defined domains.

The Web Content Mining Technique is broadly classified into four types .They are Unstructured Data, Structured Data, Semi Structured Data and Multimedia. The Semi structured is further classified into Top down Extraction, using OEM and Web Data Extraction Language. The Top Down Extraction is used to extracts multipart objects from a set of web sources and transforms them into less complex objects until particular object has been extracted. In object Exchange Model, the appropriate information is extracted from semi structured data and are inserted into a group of useful information and stored in OEM. In Web data Extraction language, it transforms web data to structured data and delivers to end users. It stores data into form of tables.

Mozenda is Software as a Service (SaaS) that allows users of all types to easily and affordably extract and manage web data. With Mozenda, users can have a set of connections agents that routinely extract data, store data, and publish data to various destinations. Screen Scrapper

A screen scrapper tool can be used to scrap. By using this tool we can able to navigate to any web site that we are looking to extract data. By Using Screen Scrappers we can easily form the data which has a common formats well-matched with databases, spread sheets, etc.

Text Web scraper is used to scrap a data from a target website in text format. The information can be used for a blend of purposes depending on the needs of the user. The text web scraper can be used to get all the text on the website or specified text. They are more efficient and fast.

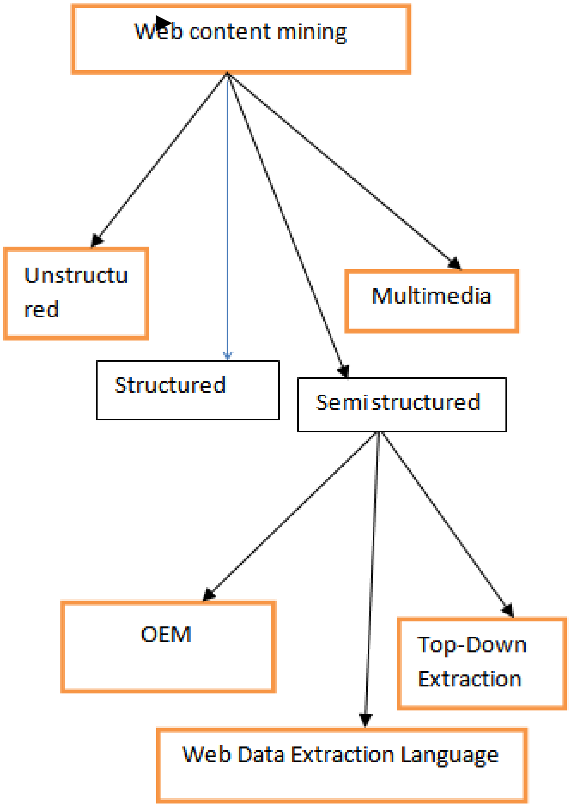

Fig 3: Web Content Mining Techniques

Automation Anywhere 6.1 (AA): AA is a Web data extraction tool which 4is used to retrieve web data. Information drive today's business and the internet is a source of power of information. Most business depend on the web to gather data that is crucial to their decision making processes Automation Anywhere can help easily automate data extraction Automation Anywhere intelligently extracts information. Running on SMART Automation knowledge Web Info Extractor (WIE): This is a tool for data mining, extracting Web content, and Web content analysis. WIE can extract structured or unstructured data from Web page, reform into home file or save to database, place into Web server.

V IMPLEMENTATION OF CONVERSION

Data Analysis

When the schema transformation is accomplished, the request data can be altered into the target mode according to its matching association and the essential procedure is divided into 3 steps:

(1) Read the document data and get format data

(2) Translate from the unique data to the target model of data according to the equivalent association between the data

(3) Mark the second step of the data to the MySQL database.

Program Implementation

The following are module implemented by the translation agenda.

(1) Doc class represents a data model class of JSON data and other data models

(2) Update Action class belongs to the database connection class

(3) Read class is the core class used to call other classes and complete all the steps of a file.

(4) The Account class is a target model class, where setter and getter represent the get and set functions for all attributes.

(5) FileOut class is the picture file output class, which converts binary data into JPG images.

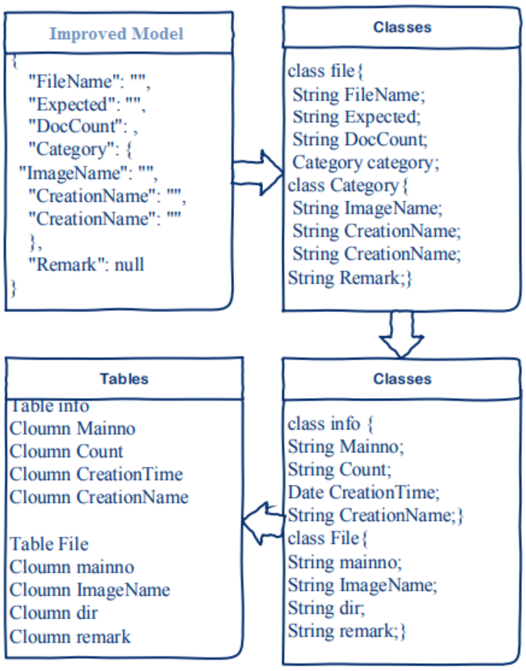

Fig 4. Improved Model Transformation Process

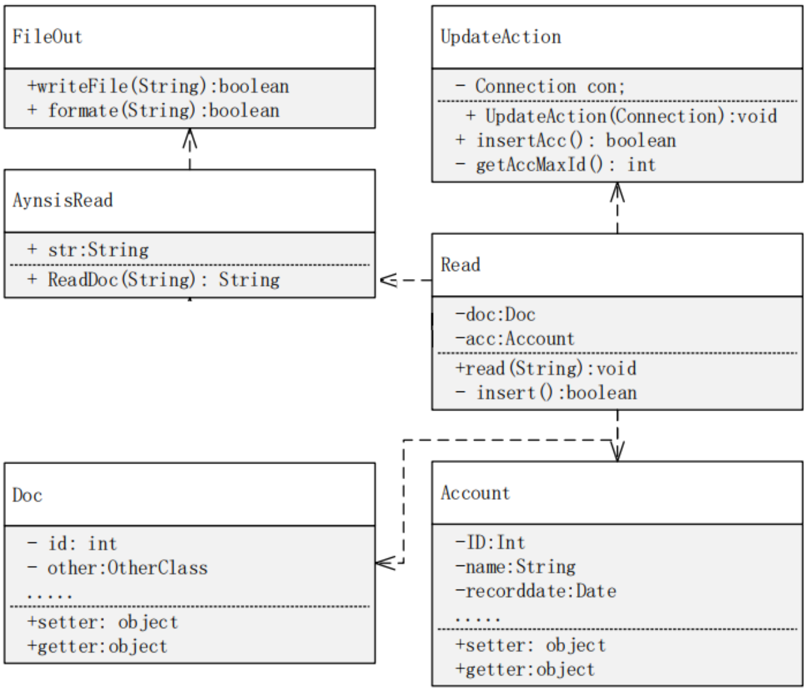

Fig 5. Program Implementation of Improved Model

VI CONCLUSION

The Web Mining Tools are used in order to take out the information and provide the outcome with best solution. The mining tools play a vital role in order to perform the extraction. Various tools are compared with the proposed model this tool has its own merit. Today's researchers concentrate on World Wide Web and each data on the web is superfluous and as day passes, additional information is placed in the server. This takes a lot of time as well as the data may not be appropriate and also the data may contain noise and outliers. To mine the proper data, the simple solution is to focus on the tools and we have to improve the time complexity and have to acquire the information or data from the contents of web page.

REFERENCES

[1] Han J, Kamber M, "Data Mining Concepts and Techniques", second edition,Morgan Kaufmann Publishers,2006,Pp.628-648.
[2] Nimgaonkar S. and Duppala, S. 2012. A Survey on Web Content Mining and extraction of Structured and Semi structured data, IJCA Journal
[3] DarshanaNavadiya,Roshni Patel", Web Content Mining Techniques-A Comprehensive Survey", International Journal Of Engineering Research & Technology(IJERT),VOL.I issue 10,December-2012.
[4] GovindMurariUpadhyay,KanikaDhingraIITM,Janakpuri,NewDelhi,India,"International Journal Of Advanced Research in Computer Science and Software Engineering " VOL.3,issue II November 2013,ISSN:2277128x.
[5] NikiR.KapadiaKanu Patel MehulC.Parikh,"Partitioning Based Web Content Mining", International Journal Of Engineering and Technology(IJERT)VOL.I Issue 3ISSN:2278-0181.
[6] DunHam,M.H.2003 Data Mining Introductory & Advanced Topics, Pearson Education.
[7] InAmadar ,S.A and Shinde,G.N 2010.An Agent Based Intelligent Search Engine System for WebMining ,International Journal on Computer Science and Engineering ,VOL 0.2,No. 3
[8] Automation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013.
[9] Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific PublishingCompany (2008).
[10] S.Balan,"A Study of Various Techniques of Web Content Mining Research Issues and Tools", International Journal of Innovative Research & Studies VOL.2 Issue 5 ISSN 2319-9725,May -2013.
[11] Michael Friendly "Milestones in the history of thematic cartography, statistical graphics, and data visualization"(2009)
[12] Niki R. Kapadia ,Kinjal Patel "web content mining techniques – a comprehensive survey",ijreas, Feb 2012.
[13] Kshitija Pol, Nita Patil, ShreyaPatankar, Chhaya Das, A Survey on Web Content Mining and extraction of Structured and Semi structured data, Heidelberg 2005
[14] M.Kitsuregawaet.al(Eds):WISE 2005,LNCS 3806,pp,763,2005,Springer- Verlag-Berlin
[15] Katharina Kaiser and Silvia Miksch(2007)
[16] Karanbirsingh and RichaSapra:An approach for Information Retrieval: Kay's Algorithm, International Journal of Emerging Trends and Technology in computer Science,Vol 3,Issue 2,ISSN 2278-6856,April 2014.
[17] AninaMadani et al, Semi Structured Documents Mining : A Review and Comparison, 17th International Conference in Knowledge Based and Intelligent Information and Engineering System, ELSEVIER,2013.
[18] Abiteboul.S, "Querying Semi-Structured Data" Database theory -ICDT ,6th International Conference ,Delphi Greece, January 8-10,Proceedings,Pages-1-18,1997.
[19] Shiren Ye and Tat Sengchua, "Learning object models from semi structured Web Documents", IEEE Transactions on Knowledge and Data Engineering, Vol 18, No.3, Mar 2006.

[20] S.K.Jayanthi and S. Prema, “Word Sense Disambiguation in Web Content Mining Using Brill‘s Tagger Technique”, International Journal of Computer and Electrical Engineering, Vol. 3, No. 3, June 2011.
[28] Eero Hyvonen, ¨ Mirva Salminen, Miikka Junnila, Suvi Kettula, “A Content Creation Process for the Semantic Web” Helsinki Institute for Information Technology (HIIT), University of Elsinki, (2004)

[21] Kamlesh Patidar, Preetesh Purohit, Kapil Sharma, “Web content Mining using Database Approach and Multilevel Data Tracking Methodology for Digital Library”, IJCSt Vol. 2, Issue 1, ISSN : 2229-4333 (Print) | ISSN : 0976-8491 (Online), March 2011.

[22] G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V. Uma, K. Sarukesi “Relevance Ranking and Evaluation of Search Results through Web Content Mining”, Proceedings of the International Multi Conference of Engineers and Computer Scientists VOL I, March 2012, Hong Kong.

[23] Sekhar Babu Boddu “ELIMINATE THE NOISY DATA FROM WEB PAGES USING DATA MINING TECHNIQUES”, GESJ: Computer Science and Telecommunications 2013|No.2 (38), ISSN 1512-1232.

[24] Shohreh Ajoudanian, and Mohammad Davarpanah Jazi, “Deep Web Content Mining” World Academy of Science, Engineering and Technology Vol: 3 2009-01-27.
Shaheen Parveen and Ajay Kushwaha, “AN APPROACH OF DEEP WEB MINING FOR DATA EXTRACTION”,INTERNATIONAL JOURNAL OF ENGINEERING SCIENCE & ADVANCED TECHNOLOGY,ISSN: 2250–3676,Volume-2, Issue-6, 1653 – 1656 ,Nov-Dec 2012

[25] Faustina Johnson and Santhosh Kumar “Web content mining using Genetic Algorithm”, ICAC3, CCIS 361,pp.82-93 2013, Springer.

[26] Yeye He, Dong Xin, Venkatesh Ganti , Sriram Rajaraman , Nirav Shah,” Crawling Deep Web Entity Pages”, GoogleInc, WSDM’13, February 4–8, 2013, Rome, Italy.

[27] Fang Doll, Mengchi Liu, Yifeng Li, “Automatic Extraction of Semi Structured Web Data”, International Journal of Database Theory and Application Vol. 6, No. 4, August, 2013.

[28] Rajashree Shettar, Dr. Shobha,” Survey on Mining in Semi structured Data”, IJCSNS, Vol 7, No 8, Aug 2007.

[29] Svetlozar Nestorov, Serge Abiteboul, Rajeev Motwani,” Extracting Schema for Semi structured Data”, Stanford University.

[30] Naveen Ahish and Craig A. Knoblock “Wrapper Generation For Semi Structured Internet Sources”, SIGMOD Record, Vol. 26, No. 4, December 1997.