ISSN: 2321-7782 (Online)
Volume 3, Issue 6, June 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Online Shop Rating System for Business Intelligence

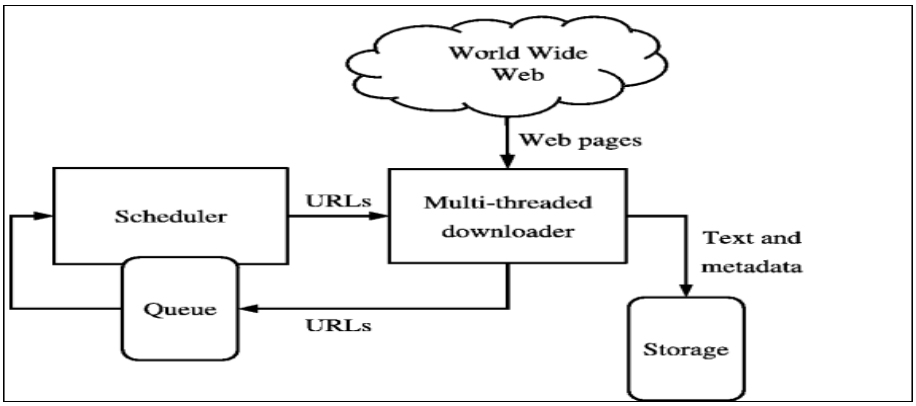
Aneesa K.P
PG scholar
MEA Engineering College
Perinthalmanna, India

Abstract: ***This is the era of revolution in the field of media especially in information technology sector, what ever information we want or solution we need the internet is there always extending its helping hand to us. There are plenty of information on internet, some of them are informative and others are non-informative. In today's search engines play a vital rolein retrieving and organizing relevant data for various purposes. A web crawler is a program, which automatically traverses the web by downloading documents and following links from page to page. Web Crawler is the main component of search engine. This paper is implemented on the data mining techniques and methods for acquiring new knowledge from data collected by online shops. The main goal of the research is to reveal the high potential of data mining applications for online shop management.***

Keywords: Web crawling, Rating, Online shops, Weightage, Information extraction

I. INTRODUCTION

The World Wide Web contains a huge collection of online resources and information where every second, new piece of information is added. As more information becomes available on the Web. Thus to find relevant information on WWW is very difficult task. . A web crawler is a program, which automatically traverses the web by downloading documents and following links from page to page. . It continuously downloads pages from WWW. These pages are indexed and stored in database. Web crawler, also called web robot or spider.



Most business applications depend on web to collect information that is crucial for decision making process. Web pages are usually generated for visualization not for data exchange. Information extraction is the task of finding structured information from unstructured or semi-structured text. It is an important task in text mining and has been extensively studied in various research communities including natural language processing, information retrieval and Web mining. It has a wide range of applications in domains such as biomedical literature mining and business intelligence. The World Wide Web has created a challenging arena for e-commerce: with a webshop, products and services can be offered to online customers. In this work, a methodology has been introduced as a decision support tool to the consumers in the Internet business. Online shops today are

operating in a very complex and highly competitive environment. The main challenge for modern shops is to deeply analyze their performance, to identify their uniqueness and to build a strategy for further development and future actions.

II. INFORMATION EXTRACTION SYSTEMS

Traditionally, the manually concepted information extraction tools are used. Manually-constructed IE systems means that users program a wrapper for each Web site by hand using general programming languages such as Perl or by using special-designed languages. So the manual extraction from semistructured data is a very difficult task. Some manually extracted information systems are TSIMMIS, Minerva, Web-OQL, W4F and XWRAP etc.[1]

TSIMMIS is one of the first approaches that give a framework for manual building of Web wrappers. In TSIMMIS each command is of the form: [*variables*, *source*, *pattern*], where *source* specifies the input text to be considered, *pattern* specifies how to find the text of interest within the source, and *variables* are a list of variables that hold the extracted results. The special symbol * in the pattern means discard and # means save the variable. The output has Object Exchange Model.

WebOQL is a functional language that can be used as query language for the Web, for semistructured data and for website restructuring. The main data structure provided by the WebOQL is the hypertree. Hypertrees are arc-labeled ordered trees which can be used to model a relational table, a Bibtex file, a directory hierarchy, etc. The tree structure is similar to the DOM tree structure where arcs are labeled with records with three attributes Tag, Source, Text, corresponding to tag name, the piece of HTML code, and the text excluding markup, respectively.

W4F is a Java toolkit to generate Web wrappers. The wrapper development process consists of three independent layers: *retrieval*, *extraction* and *mapping* layers. In the retrieval layer, a to be processed document is retrieved cleaned and then fed to an HTML parser that constructs a parse tree following the Document Object Model (DOM). In the extraction layer, extraction rules are applied on the parse tree to extract information and then store them into the W4F internal format called Nested String List (NSL). In the mapping layer, the NSL structures are exported to the upper-level application according to mapping rules.

III. LITERATURE SURVEY

As the growth of the World Wide Web exceeded all expectations, the research on web mining is growing more and more. The process of information extraction from Web is both interesting and challenging, which could be helpful in Web Searching, Information Retrieval and Web Mining. Web pages on many Web sites are produced dynamically as structural records. The web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services.

Chia-Hui Chang and Shao-Chen Lui [3], propose IEPAD, a system that automatically discovers extraction rules from Web pages. The system can automatically identify record boundary by repeated pattern mining and multiple sequence alignment. This system that attempts to to generate repetitive patterns from unlabeled web pages. The system IEPAD mainly includes three components ie, extraction rule generator, pattern viewer and extractor module. An extraction rule generator which accepts an input Web page. Extraction rule generator includes a token translator, PAT tree constructor, pattern discoverer, a pattern validator and an extraction rule composer A pattern viewer is a graphical user interface, which shows repetitive patterns discovered. An extractor module extracts desired information from similar Web pages according to the extraction rule chosen by the user. The discovery of repeated patterns is realized through a data structure called PAT trees and string alignment techniques. IEPAD exploits the fact that if a web page contains multiple data records, they are often rendered regularly using the same template for good visualization. IEPAD extracts records using wrapper induction method.

J. Wang and F.H. Lochovsky[5], proposes a data extraction method as DeLA. This technique concentrates on pages that querying back end database using complex search forms other than using keywords. DeLa automatically extract data from web site and assigns meaningful labels to data. It is wrapper generation process. It's works two steps, one is Data-rich section

extraction algorithm and another is pattern extractor algorithm. Data-rich Section Extraction algorithm (DSE) is designed to extract data-rich sections from the Web pages by comparing the DOM trees for two Web pages (from the same Web site), and discarding nodes with identical sub-trees. Pattern extractor is used to discover continuously repeated (C-repeated) patterns using suffix trees. By retaining the last occurrence for each discovered pattern, it discover new repeated patterns from the new sequence iteratively, forming nested structure. Finally, labels are assigned to the columns of the data table by four heuristics, including element labels in the search form or tables of the page and maximal-prefix and maximal-suffix shared by all cells of the column.

Mohammed Kayed and Chia-Hui Chang [6], proposes FivaTech, which is a page-level web data extraction technique. It mainly consists of two modules:one is make fixed/variant pattern tree and other is detect the schema. Fixed/variant module means that DOM trees of web pages as input and merges all DOM trees into a structure. In the second module template and schema are detected from fixed/variant pattern tree.In first module all nodes of input DOM trees into a matrix form. This module can be divided into four sub modules. They are Peer node recognition, Multiple string alignment, Pattern mining, Optional node merging. Nodes which have same tag name but different functions are called peer nodes. Peer nodes are denoted using same symbol inorder to facilitate string alignment. Pattern mining on aligned string will remove extra occurrences of discovered pattern. Optional node merging step recognizes optional nodes, the nodes which are which disappears in some column of the matrix. Schema detection module detects structure of the website ie, identifying the schema and defining the template.

Chaw Su Win, Mie Mie Su Thwin[7], proposes a new method for information extraction from web pages. It mainly consists of two steps: web page segmentation and informative content extraction. So this paper Effective Visual Block Extractor (EVBE) Algorithm for web page segmentation. And it also proposes Effective Informative Content Extractor (EIFCE) Algorithm for web informative content extraction. A web page structure and layout varies depending on different content types and the presentation style designed by the web developer. Thereby informative content positions of the web pages differ in variety of websites. In this case for effective web page segmentation have three steps: cleaning the web page, DOM tree construction and visual block extraction are carried out. Cleaning the web page means that the web pages are not well formed structure they contain contain invalid tag structure such as there is an opening tag with no corresponding closing tag and vice versa. Therefore these invalid tag structures are needed to be cleaned before processing them. The DOM tree construction means that presents an HTML document as a tree-structure. Each HTML page corresponds to a DOM tree where tags are internal nodes and the actual text, images or hyperlinks are the leaf nodes. Visual block extraction approach for the extraction of semantic blocks from web page by applying EVBE Algorithm. The Algorithm uses the DOM structure of the input web page and the visual properties of each DOM node for effective extraction of semantic blocks from web page. The Algorithm recursively traverses the input DOM tree in pre-order manner and returns the Block Tree and the General Parameters of each block as its output. After segmenting the web page into semantic blocks correctly, the Informative Content Block of the web page can be extracted effectively. The Algorithm applies Choose Candidate Blocks function for choosing Candidate Blocks for the Informative Content Block. And then it applies the Find Informative Block method to detect the Informative Content Block from the Candidate Blocks. Finally it returns the Informative Content Block of the web page as result. Effective Informative content Extractor algorithm has two steps: choose candidate block extraction function and finding informative content block.

Arvind Arasu and Hector Garcia-Molina[8], presented an effective formulation for the problem of data extraction from Web pages. Broadly, EXALG works in two stages. In the first stage (ECGM), it discovers sets of tokens associated with the same type constructor in the (unknown) template used to create the input pages. In the second stage (Analysis), it uses the above sets to deduce the template. The deduced template is then used to extract the values encoded in the pages. The input of EXALG is a set of pages created from the unknown template *T* and the values to be extracted. EXALG deduces the template *T* and uses it to extract the set of values from the encoded pages as an output. ECGM stage computes equivalence classes. i.e, set of tokens

having same frequency of occurrence in every page. This is performed by FindEquiv sub module. EXALG only considers equivalence classes that are large and contain tokens which occur in large number of pages. Such equivalence classes are called Large and Frequently Occurring Equivalence Classes(LFEQs). Sub module HandInv detects and removes invalid LFEQs. DiffFormat sub module differentiates roles of non tag tokens using context in which they occur. Occurrence path of a page-token is path from root to page-token in the parse tree representing the page. Analysis stage constructs a template using LFEQs. For this EXALG first considers root LFEQs whose tokens occur exactly once in every input page. Position between two consecutive tokens is empty if they occur contiguously otherwise it is non empty. EXALG will determine tokens of LFEQs which are non empty. It constructs output template T' by generating a mapping from each type constructor in S' to ordered set of strings.

IV. PROPOSED WORK

Word of mouth is one of the oldest and most effective forms of marketing, and its influence is one on the rise as traditional marketing becomes more expensive and less effective. In the past, word of mouth was limited to person to person communication. Today, the viral nature of internet allows conversations to quickly spread around the world and millions of people. This fact, coupled with a growing distrust for traditional advertising, has made online ratings and rankings one of the most powerful factors influencing consumer purchasing decision. Ratings are important to people, especially for building up trust among business partners. The traditional offline rating systems have always been and still are word of mouth and gossip.

As a huge data source the internet contains a large number of valuable information, and the data of information is usually in the form of semi-structured in HTML web pages. Most of the web resources are in the form of Hypertext Markup Language (HTML) documents, which are viewed by web browsers. Web mining helps to make the process of finding the needed information from a huge data like the web in an effective and effective way. It consists of various actions like, analytics, databases, processing, information retrieval, multimedia, etc. Web Information Extraction (IE) tools that extract information and knowledge from the Web pages and transfer it into a meaning and useful structures for further analysis will become a great necessity. There are a lot of approaches have been developed in the area of Web IE which concerns with how to harvest useful information for any further analysis from web pages. Web IE aims at extracting text from online documents that are semi-structured and usually generated automatically by a server-side application program. In Web IE, usually machine learning and pattern mining techniques are applied to exploit the syntactical patterns or layout structures of the template-based documents.

Websites are set of related web pages typically served from a single domain. Each Web site contains a home page, which is the first document users see when they enter the site. The site might also contain additional documents and files. Each site is owned and managed by an individual, company or organization. Each of the webpage contain various data, hyperlinks images etc. Internet web pages typically contain a large amount of non-informative content such as advertisements, search and filtering panel, headers, footers, navigation links, and copyright notices, etc. Most clients and end-users search for only the informative content and the need of Informative Content Extraction from web pages becomes evident.

a) System Architecture

This work proposes a rating system for the shops based on the some features or parameters. .A crawler starts from a set of seed pages (URLs) and then uses the links within them to fetch other pages. The links in these pages are, in turn, extracted and the corresponding pages are visited. The process repeats until a sufficient number of pages are visited or some other objective is achieved. A crawler can visit many sites to collect the information that can be analyzed and mined in a central location. Content extractions are mainly based on the user's interesting parameters or patterns. These parameters are arranged on the each of the category. Different categories have different common parameters. These content extraction are based on the corresponding shops in each category. In rating each and every parameters value have a score. Weightage calculation is the summation of the score values. In ranking re-ordered the results to the user.

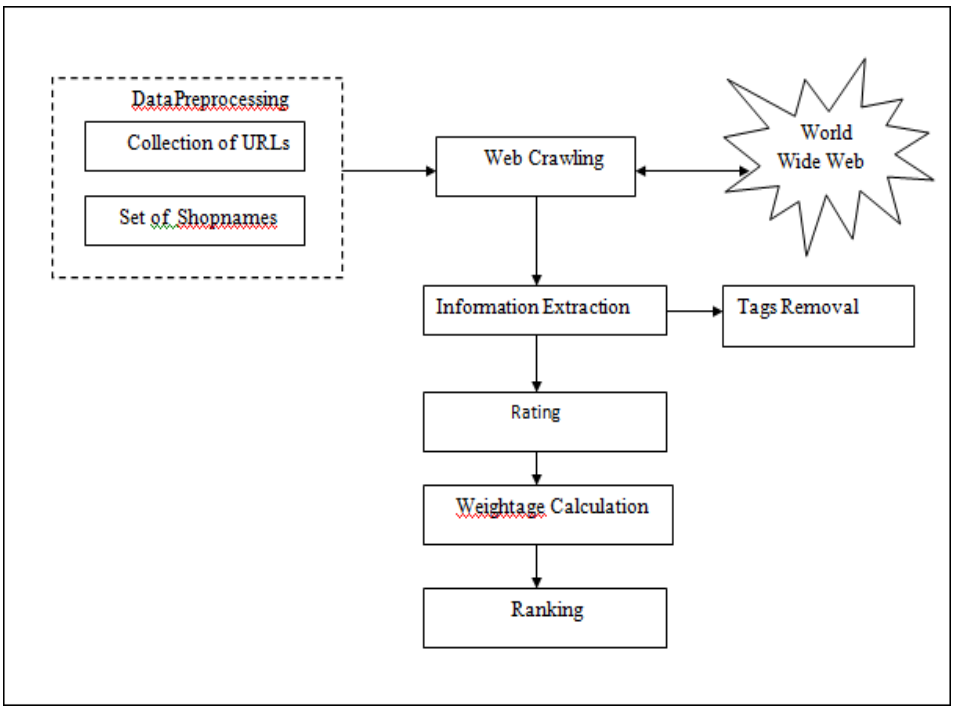


Fig: System architecture

V. CONCLUSION

World Wide Web continuously introduces new capabilities and attracts many people. Web crawling is the process by which we gather pages from the Web, in order to index them and support a search engine. The objective of crawling is to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them. It is the skill of surfer to extract the information he/she wants without consuming much time. In my work, I have tried to develop an application which helps the online shoppers to easily access the shops and its details which the shopper is in search. My work helps the customers find out the highly rated shops with their special features offers and facilities. On each page of my work there won't find any unnecessary information.

ACKNOWLEDGMENTS

We take this opportunity to express our gratitude to all who encouraged us to complete this work. We would like to express our deep sense of gratitude to MEA engineering college, Perinthalmanna to support our work. And also wish to express heartfelt thanks to the anonymous reviewers for their all contribution to improving the quality of this paper.

References

1. C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaala, "A survey of web information extraction systems," Knowledge and Data Engineering, IEEE Transactions on, vol. 18, no. 10, pp. 1411–1428, 2006
2. R. R. UTKALUNIVERSITY, "Integration of web mining and web crawler: Relevance and state of art," Integration, vol. 2, no. 03, pp. 772–776, 2010.
3. C.-H. Chang and S.-C. Lui, "Iepad: information extraction based on pattern discovery," in Proceedings of the 10th international conference on World Wide Web. ACM, 2001, pp. 681–688.
4. K. Devika and S. Surendran, "An overview of web data extraction techniques," International Journal of Scientific Engineering, 2013.
5. J. Wang and F. H. Lochovsky, "Data-rich section extraction from html pages," in Web Information Systems Engineering, 2002. WISE 2002. Proceedings of the Third International Conference on. IEEE, 2002, pp. 313–322.
6. M. Kayed and C.-H. Chang, "Fivatech: Page-level web data extraction from template pages," Knowledge and Data Engineering, IEEE Transactions on, vol. 22, no. 2, pp. 249–263, 2010.
7. C. S. Win and M. M. S. Thwin, "Informative content extraction by using eifce [effective informative content extractor]," International Journal of Scientific & Technology Research, vol. 2, no. 6, pp. 136–144, 2013.
8. A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM, 2003, pp. 337–348.

