Extraction of Most Relevant Data from Deep Web Mining

Ashok P, V. Hariharan, R. Lavanya, R. R. Prianka

Abstract- Extraction of web content from the deep web page is the tough task to retrieve the relevant data because they are web page programming language dependent. The challenges of such web page extraction are increases every day due to expanding of huge web database, which makes the researchers to concentrate on deep web mining. Whenever user submits a query into search engine, it retrieves the list of best matching web page with short summary of notes such as title, some text from specific site. But retrieved information from web database is locked as deep web (Hidden Web or Invisible Web) on web page. In this paper, we proposed ontological technique with WordNet to extract the data records from the deep web pages. This technique discovers best matching words, eliminates unnecessary tags and able to extract large variety of data records with different structures.

Keyword: Ontology, Deep web, WordNet, Web Mining

I. INTRODUCTION

Deep web page started at 1994 known as Hidden Web and later it was renamed as Deep Web in 2001. Web Database contains huge volume of data that retrieve the information according to user's queries. Most of retrieved information is in the form of dynamic page. Due to this nature, generated information forms Hidden web page that is usually enwrapped in HTML page as data record and it is hard to index by search engines. Generally web page contains some non related items such as navigation, decoration, contact information, fonts, and interaction. In this paper, we proposed three-way filtering concept to eliminate the unwanted web items and able to provide only users related content of information. Our results shows better performance when compare to other extraction methods.

II. WORDNET

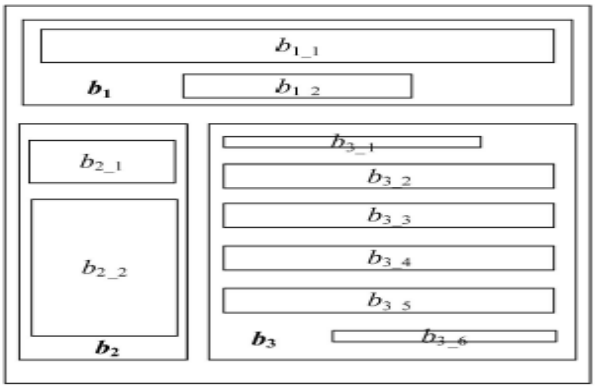
In 1998, a new lexical database called WordNet was developed for finding the semantic matching of English words. WordNet used to manage and navigate the entity component on web page. It represents synsets by means of conceptual semantic and lexical relationship between words. It classifies English words into numerous groups, such as hypernyms, synonyms, and antonyms. In general, semantic matching of words can be divided into four categories.

The initial category measure the similarity of words based on two terms as length of the path between the terms and position of the terms.

In the next category, the similarity is considered by examining the difference in content of the two terms using a probabilistic function. For the third type, similarity of words is measured using the two terms as a function of their properties (e.g. gloss overlap) or based on their relationship with other similar terms in the taxonomy. Finally, the last category measures similarity of words by combining the methods

III. STRUCTURE OF DEEP WEB PAGE

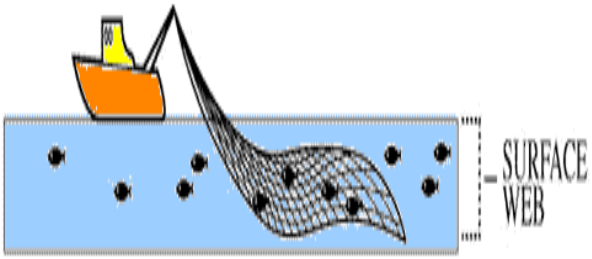
The Visual Information about the structure of deep web page is represented with help of VIPS algorithm. The VIPS algorithm used to transform the deep web page into visual blocks.



The above visual block represents the VIPS transformed deep web page structure that is segmentation of web pages. The root block represents the whole page and each block in the tree corresponds to a rectangular region on the Web page. The leaf blocks cannot be segmented more, and they characterize the minimum semantic units, such as continuous texts or images.

IV. SURFACE WEB VS DEEP WEB

4.1 Surface Web



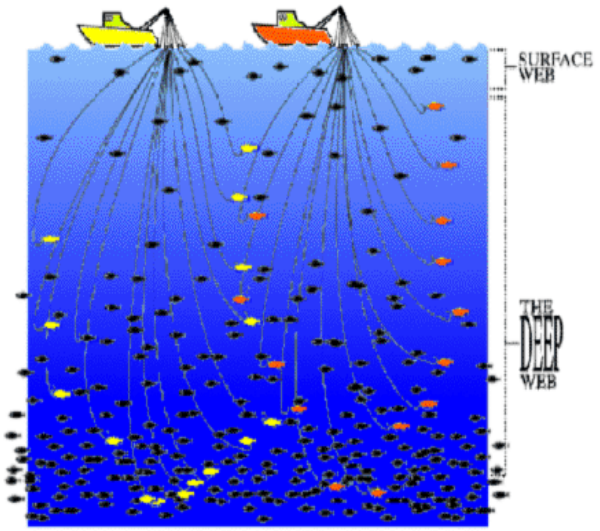
Manuscript Received on December 2014.

P. Ashok, Department of Information Technology, Veltech Hightech Dr. R. R. S. R. Engineering College, Avadi, Chennai, India.
V. Hariharan, Department of Information Technology, Veltech Hightech Dr. R. R. S. R. Engineering College, Avadi, Chennai, India.
R. Lavanya, Department of CSE, Veltech Multitech Dr. R. R. S. R. Engineering College, Avadi, Chennai, India.
R. R. Prianka, Department of Information Technology, Veltech Hightech Dr. R. R. S. R. Engineering College, Avadi, Chennai, India.

The surface web is also known as clearnet which is a part of www and it is indexable by conventional search engine, which consist of loosely speaking, interlinked HTML pages. Once user requests the required information by searching on web, surface web identifies only the content what appears on the surface and remaining data are hidden deeper. A graphical depiction of the above diagram represents the limitations of the typical search engine. The required content searched by users is identified only what appears on the surface and the harvest is comparatively indiscriminate. There is tremendous value that resides deeper than this surface content. So, surface web search is not suitable to web search than deep web search.

4.2 Deep Web Page

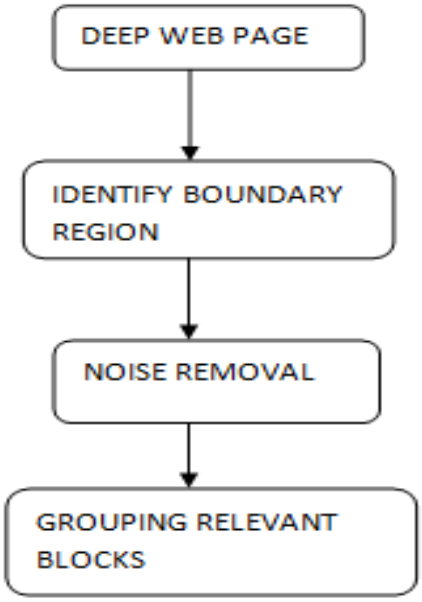
The Structure of Deep Web Page is based on huge graphs twisted by centralized crawlers and indexers. The deep Web is qualitatively dissimilar from the surface Web pages, it store their content in searchable databases and provide dynamic results in response to a direct users request. Typical, deep Web page sites receive fifty per cent greater than surface sites in monthly traffic and are more highly linked to than surface sites. The deep Web is the major rising type of new information on the Internet and its sites are tending to be narrower, with deeper content, than conventional surface sites. The Total quality content of the deep Web is 1,000 to 2,000 times greater than that of the surface Web.



The above picture represents, in a non-scientific way, the enhanced outcome that can be obtained by BrightPlanet technology. By initial identifying where the appropriate searchable databases reside, a directed query can then be placed to each of these sources at the same time to produce only the results preferred with pinpoint accuracy.

V. DEEP WEB DATA EXTRACTION

The web pages which are not indexed by the search engines are called deep web pages, example-dynamic web pages. The data records which are located in the deep web are semantically related and also share a common tree structure. Wrappers designed with ontological technique improve the accuracy of the deep web data extraction. If domain independent wrapper is designed then a vast amount of data can be extracted. An Ontological wrapper can be designed to extract data from the deep web. The main steps for designing an ontological wrapper are (i)Deep web pages needs to be parsed (ii)The unwanted components needs to be filtered by using suitable filtering component. WorldNet can be used the semantically related components can be used to extract the relevant components from the deep web. Using ontological techniques with the wrapper for web data extraction makes the wrapper more robust. The sizes of the data records in the deep web are three times larger than a normal web page. The earlier methods which were used for web data extraction are a semiautomatic method XWRAP and automatic method ROADRUNNER, all are structured based methods. For extracting the data from deep web pages the boundary needs to be identified first and then data has to be extracted. As a preprocessing step the noise needs to be eliminated. The relevant blocks grouped together. The grouping is done based on the semantics. Then relevant data item is extracted



5.1 Identifying Boundary

The fact to be considered while identifying the data region boundary; there could be blocks that must not belong to further data record and annotation about data records.

5.2 Noise Removal

Generally these blocks are aligned at the top or bottom of the web page and this type of blocks do not contain any data records. This phase does not guarantee about the elimination of complete noise blocks.

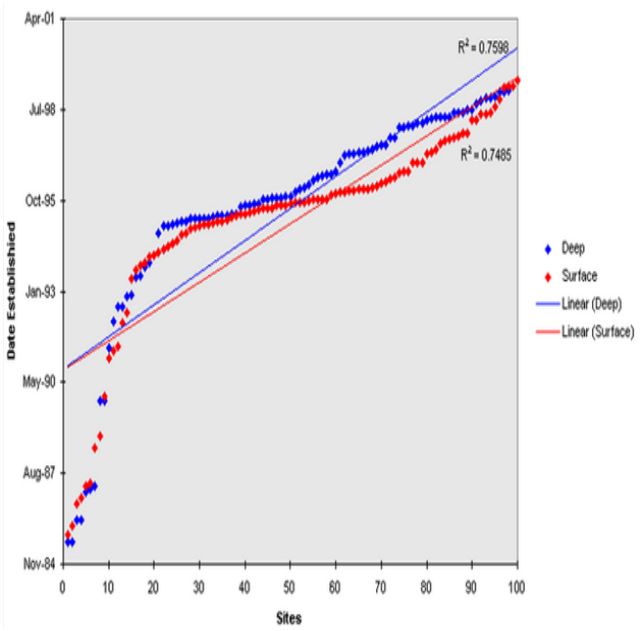
5.3 Grouping Relevant Data

In grouping relevant data, wrapper uses an ontological technique to check the similarity of data records.

This technique is able to filter out irrelevant data region such as menus which determines the layout of a web page and also able to reduce the candidates for data extraction, hence results in higher accuracy in data extraction.

VI. PERFORMANCE METRICS

6.1 Growth Rate of Deep Web vs Surface Web.



Time-series analysis used to measure the growth rate between surface web and deep web page sites. The top method to test for concrete growth is a time series analysis. BrightPlanet plans to institute such tracking mechanisms to achieve better growth estimates in the future. Thus, it shows that the deep web growing faster than the surface web search with short period of time.

6.2 Higher Quality on Deep Web Page

Query	Surface Web			Deep Web		
	Total	"Quality"	Yield	Total	"Quality"	Yield
Agriculture	400	20	5.0%	300	42	14.0%
Medicine	500	23	4.6%	400	50	12.5%
Finance	350	18	5.1%	600	75	12.5%
Science	700	30	4.3%	700	80	11.4%
Law	260	12	4.6%	320	38	11.9%
TOTAL	2,210	103	4.7%	2,320	285	12.3%

The complete number of results in above table shows that deep web sites tend to return valid conclusion that quality is many times greater for the deep Web than for the surface Web.

VII. CONCLUSION

Our projected ontological technique, extracts the data records from web page with varying structures efficiently. The ontological technique could also shrink the number of potential data regions that is used for data extraction and this will shorten the time and increase the accuracy in identifying the accurate data region to be extracted and it thus provides more flexibility to extraction the complicated data records from web page.

REFERENCE

1. MICHAEL K. BERGMAN, "The Deep Web: Surfacing Hidden Value," The journal of Electronic publishing.
2. Christiane Fellbaum, "WordNet: An Electronic Lexical Database," The MIT Press, Cambridge, MA, 1998.
3. Munindar P. Singh • *North Carolina State University* • *singh@ncsu.edu*
4. Wei Liu, Xiaofeng Meng, Member, IEEE, and Weiyi Meng, Member, IEEE "ViDE: A Vision-Based Approach for Deep Web Data Extraction "
5. B.Aysha Banu, M.Chitra," A Novel Ensemble Vision Based Deep Web Data Extraction Technique for Web Mining Applications", "IEEE Intemational Conference on Advanced Communication Control and Computing Technologies (ICACCCT) ",2012
6. D. Cai, S. Yu, J. Wen, and W. Ma, "Extracting Content Structure for Web Pages Based on Visual Representation," Proc. Asia Pacific Web Conf. (APWeb), pp. 406-417, 2003.
7. Bergman, Michael K,"White Paper: The Deep Web: Surfacing Hidden Value", Volume 7, Issue 1: *Taking License*, August, 2001 DOI: http://dx.doi.org/10.3998/3336451.0007.104
8. Jer Lang Hong "Deep Web Data Extraction" School of IT, Monash University
9. "Sasikala.D1, Selva Kumar.G2" Extraction of Deep Web Contents" International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.1, Jan-Feb 2012 pp-528-533