Australian Journal of Basic and Applied Sciences, 9(11) May 2015, Pages: 451-460

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Word-Location Based Indexing Using Sequential Pattern in Deep Web Mining for Query Processing

[1]Vijay R. and [2]Dr. Prasadh K.

[1]*Research Scholar, M.S University, Tamilnadu, India,*
[2]*Mookambika Technical Campus, Eranakulam, Kerala, India,*

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form 26 January 2015
Accepted 27 February 2015
Available online 19 April 2015

Keywords:
Word-Location based Indexing, Sequential Pattern, Meta-location, Cosine Similarity, Prefix Deep Word-Location Mining Process.

ABSTRACT

The rapid growth and massive quantities of data on the internet have increased the importance and complexity of information retrieval through deep web mining. The deep web mining creates new challenges for exact information retrieval as the amount of information on the web is growing rapidly. Research works conducted for automatically identifying and extracting the common lists on the Web using HyLiEn (Hybrid approach for automatic List discovery and Extraction on the Web), uses the information from visual alignment of list items. HyLiEn approach extracts general lists from web documents but not effective in indexing the web based on lists and tables. A novel existing approach Visually Extracting Data Records (VEDR) from the deep web identifies each data record individually but the queries are not answered clearly. To develop the effective indexing on the web, Word-Location based Indexing Using Sequential Pattern (WLI-SP) is employed. WLI retrieve effective result through deep web mining based on the meta-location. The meta-location denotes the URL, title and header and WLI-SP used meta-location to index both the long and shorter text for the user query. WLI-SP performs the searches using the Merkle word location search tree. WLI-SP model measure the closeness of meta-location for two different documents using the cosine similarity. Sequential pattern growth in WLI-SP model effectively and efficiently pushed deep mining into prefix deep word-location mining process. Prefix deep word-location mining process mine and answer the user queries clearly with query form such as 'and', 'or', 'neither-nor', 'either-or' respectively. To measure the performance of WLI-SP model, experiments were performed on JAVA Weka tool using the factors such as relative runtime, averaged accuracy ratio, deep mining query result ratio, true positive rate, memory usage and sequential pattern based similarity measure.

© 2015 AENSI Publisher All rights reserved.

To Cite This Article: Vijay R. and Dr. Prasadh K., Word-Location Based Index Using Sequential Pattern in Deep Web Mining for Query Processing. ***Aust. J. Basic & Appl. Sci.,*** *9(11): 451-460, 2015*

INTRODUCTION

Deep web mining produces new challenges for user required information retrieval as the quantity of information on the web is increasing rapidly. The explosive growth and popularity of the world-wide web has resulted in a large amount of information sources on the internet. Conversely, due to lack of web structure information sources, access to huge collection of information has been limited to browsing and searching. Sophisticated deep web mining applications, such as judgment of shopping robots, require travel expense analyzes to deal with different data formats.

Structured-Web Threaded Crawler (SWTC), a simple algorithm that hold the problem by providing auxiliary sessions to the wrapper which optimize its output. The threaded algorithm as expressed in (Boutros R., 2013) guarantee accessing of all available detail pages within crawling scope but the testing environment fails to include additional domains with different input settings. SWTC is not effective in handling more crawling scenarios with diverse domains. Visual based wrapper extract information from HTML pages as shown in (Leng Goh, 2031) uses visual cue to remove avoidable regions. The unnecessary regions minimizes the running time as the extraction task of the wrapper only needs to consider the relevant region for extraction.

A quickly growing number of deep web databases are now turning out to be accessible via their HTML based query interfaces. The structured web-sources, query schemas are discriminative representatives of sources that took advantage of the ordinary observation. The volume of structured data on the deep web has been growing extremely large. Such data is typically returned from the back-end

Corresponding Author: Vijay, R. Research Scholar, M.S University, Tamilnadu, India,
E-mail: vijayrengitham@gmail.com

web databases in response to explicit user queries, and presented in the form of data records, which are enwrapped in query result pages.

An automated approach as illustrated in (Daiyue Weng, 2012) for web data extraction formulate the use of visual features and query terms to identify data sections and extracts data records. The numerous content and visual features of visual blocks in a data section fail to expand algorithms for annotating data items in the extracted data record. So that data items of the same attribute cannot be labeled to indicate its meaning in the database table.

Visual Architecture based web information extraction in (Oswalt Manoj, S., 2011) seeks a diverse way for deep web data extraction to mitigate the limitations of previous works by utilizing certain attractive common visual features on the deep web pages. ViDE as illustrated in (Wei Liu, Xiaofeng Meng and Weiyi Meng, 2010) can only process on deep web pages containing one data region, while there is a significant amount of failure to perform on multi data-region deep web pages. ViDE is not intended to tackle HTML dependent and programming AIP of IE, which is time consuming process.

Fully automatic algorithm as illustrated in (Roberto Panerai Velloso, 2013) uses a Tag Path Sequence (TPS) representation of the web page. The TPS consists of a sequence of symbols to represent a different tag path. The TPS searches for positions to split it in two regions where each region's alphabet do not intersect, which means that they have totally different sets of tag paths and, thus, with different regions. ODE (Ontology-assisted Data Extraction), automatically extracts the query result records from the HTML pages in (Weifeng, S.U., 2009). ODE first constructs ontology for a domain according to information matching in the middle of the query interfaces and query result pages from different web sites inside the similar domain. The constructed domain ontology is further used during the process of data extraction to identify the results of query and to mark the data values in the extracted records.

Template Extraction as shown in (Trupti, B., 2012) extracted the heterogeneous web pages. But due to huge form of web documents, there is a requirement to direct unidentified number of templates by clustering web documents for choosing a good partition method. Document Object Model (DOM) tree of HTML document find the essential paths of document but accuracy of extracted templates was not effective in cluster quality maintenance. Two similarity based mechanisms Web Data Extraction and Integration System (WDES) as expressed in (Srikantaiah, K.C., 2013) take out desired SERP and store them in the local depository for offline browsing. WDICS integrate the requested contents and enable the user to perform the planned analysis and extract the preferred information, not effective in indexing.

Algorithms for extracting templates from a huge count of web documents as shown in [12] produce heterogeneous templates. Extraction of business rules from web logs as demonstrated in [10] allow logical reasoning to draw inferences based on the logical content of the input data. Certain types of software tools used to produce wrappers are Wrapper induction (WI) and Information extraction (IE) systems where the task of a wrapper is to typically perform a pattern matching procedure.

MAximal Resemblance Data Labeling (MARDL) as allocated in (Hung-Leng Chen, 2009) surround every unlabeled data point into the corresponding suitable cluster based on the novel categorical clustering representative for pattern matching. N-Nodeset significance representative symbolize clusters of different cluster combination. MARDL demonstrate high execution efficiency, and accomplish high intra cluster relationship and low inter-cluster relationship. Several concept-based user profiling methods as illustrated in (Kopal Maheshwari, Namrata Tapaswi, 2013) are competent of deriving both the user's positive and negative preferences. The entire user profiling strategies were query-oriented, but the relationships between users fail to perform collaborative filtering.

The existing user profiles cannot be used to predict the intent of unseen queries, such that when a user submits a new query, personalization fail to get benefit from the unseen query. The possibility to identify the leakages was performed in an efficiency manner using the data allocation strategies as demonstrated in (Panagiotis Papadimitriou and Hector Garcia-Molina, 2011). The methods rely on alterations of the released data but built models do not capture leakage scenarios. Finally, also judge the choice of adding "fake" objects to the distributed set. An additional open problem is the extension of allocation strategies as they can hold only agent requests in an online fashion.

One of the hybrids methods as explained in (Fabio Fumarola, 2011) performed the task of extracting the portions of general lists from the web and uses general assumptions on the visual rendering of lists. In case of the structural representation of item, it is present in the form of both visual and structural features of web lists. The work on Hybrid approach for automatic List discovery and Extraction (HyLiEn) identified and extracts the common lists present on the web in a mechanical manner. HyLiEn uses the information on chart alignment of list items for discovery but not effective in indexing the web based on lists and tables.

In this work, Word-Location based Indexing Using Sequential Pattern (WLI-SP) retrieve the user result effectively using meta-location. The meta-location contains the URL, Title and then Header form. Sequential pattern growth in word-location based indexing efficiently and proficiently pushed deep to get the user result. The prefix deep word-

location mining process uses the depth first strategy. Prefix deep word-location mining process mine and answer the user queries clearly with different querying strategies. The experiment is performed on the different parametric factors to yield the clear query result in the deep web mining and result is compared with the existing system through graph values.

The structure of paper is as follows. In Section 1, describe the deep web mining drawbacks of the existing web based indexing on lists and tables. In Section 2, demonstrates the Word-Location based Indexing Using Sequential Pattern model for answering the user query clearly. Section 3 explains about the experimental setup with parametric factor description. Section 4 analysis the result through table and graph values and section 5 illustrates the existing works with limitations. Section 6 finally concludes with beneficial deep web result for the users query

Word-Location Based Indexing Using Sequential Pattern:

The main objective of Word-Location based Indexing using Sequential Pattern is to answer the user queries clearly with the help of data tables and data lists from web database. The user query is presented in the form of different query words.

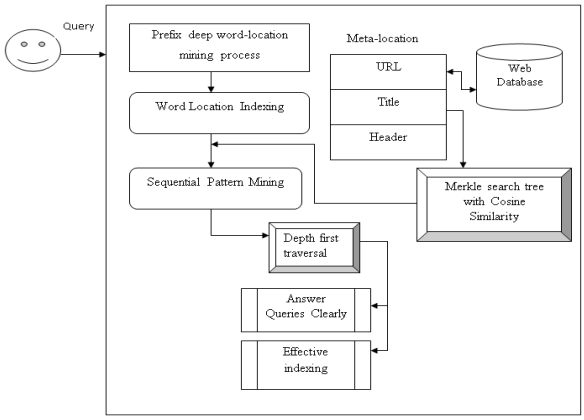

Fig. 1: System Architecture of Word-Location based Indexing Using Sequential Pattern model.

The location of the word is initially identified using the meta-location form. The meta-location comprises of the URL, Title and then Header. Based on the general structure of URL, the URL form for WLI-SP model is classified into main domain and sub domain. The title consist of the 'keywords' in the query and header consists of the different query forms. For example, the header 'h1' contains 'and', 'h2' contains 'or', 'h3' contains the 'neither-nor', 'h4' contains the 'either-or' words and continues till 'hn'. The system architecture of WLI-SP is shown in Fig 1.

Fig 1 shows the system architecture with the overall system process based on the lists and tables. The two processes involved in the proposed WLI-SP model are (i) word location indexing and (ii) sequential pattern mining. In order to extract the answer for the query provided by the user, word location based indexing uses the cosine similarity for each term from the web database. The cosine similarity in WLI-SP matches the location of word from the user query sentence and the web database while performing the deep web mining. Sequential pattern mining is used in WLI to handle shorter and longer user queries in deep mining.

Sequential Pattern Mining uses the prefix deep word-location mining process to answer the user query clearly with word-location indexing concept. The process involved in prefix deep word-location mining solves the constraints related to the user query in an efficient way. Prefix deep word-location push produce results on various normally used query constraints into sequential pattern mining. Prefix deep word-location mining is efficient in sequential pattern mining which addresses the problem of indexing using meta-location. WLI-SP model return the deep mining result not only on specific types of query constraints but also related to certain complex queries, such avg () and sum ().

Word Location Indexing:

To start with the proposed WLI-SP model, the first process, word location indexing is briefed with the help of the Merkle word location search tree as illustrated in fig 2. The idea behind word location indexing is to retrieve the relevant documents in deep web mining using specific meta-location points. Word location indexing in WLI-SP model uses the Merkle word location search tree with meta-location. The advantage of using Merkle word location search tree is that it handles and searches large collection of web documents from web database.

The Merkle search tree in WLI-SP model consists of a tree of words with set of files. In the top of a Merkle tree consist of a top search result with trusted resource. The word-location based indexing uses the meta-location to check in coherence with the trusted top result. If the Merkle

word location search tree is damaged, another tree is tried and the process continues until the program identifies one that matches with the top result. The damaged Merkle tree tries with files of different sites (i.e.,) URL for similar result. The closeness between the previously developed Merkle search tree results and newly developed search results is computed through cosine similarity

The cosine similarity between the two files is computed as given below:

$$Similairty\ (F1,F2\ldots Fn) = \frac{\sum_{f=1}^{t} WQT_f, URL_f, Title_f, header_f}{\sqrt{\sum_{f=1}^{t}(WQT_d{}^2)\sum_{f=1}^{t} URL_f, Title_f, header_f{}^2)}} \quad \text{Eqn (1)}$$

Where, WQT_f denotes the weight of Query Term'f', specified with the URL represented by URL_f with the domain and sub domain name of 'f'. The title is obtained using $Title_f$ which represents the keyword used in the query term'f' with the $header_f$ representing the different query form constrains of 'f', for t denoting the number of file 'f' terms used for similarity measurement.

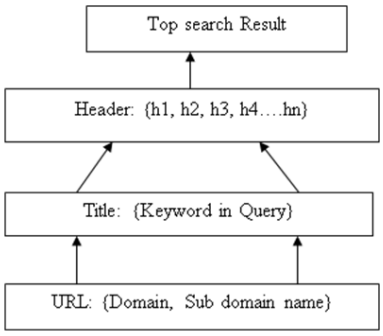

Fig. 2: Merkle word location search tree.

The similarity between the Files 'F1', 'F2' …'Fn' is measured effectively using the cosine similarity which is used to measure individually for each file with the meta-location. Each meta-location is assigned with its cosine similarity measure in deep web mining in addition to the file content. To determine the best query result to the users, a specified threshold value is chosen to easily select the best result with meta-location by ignoring the other results.

The word location indexing provides very easy result by selecting the keywords, URL and headers from the user query. To store the user query information, each file has the meta-location with certain dimension and cosine measure holds the 'fileID' and 'KeywordID'. File relevance in WLI-SP model is computed by adding the cosine similarity for the user query term. The meta-location on 'n' files holds the word-location based indexing in the form of

$$File\ ID, Keyword\ ID = \{(f1,k1),(f2,k2),(f3,K3)\ldots.(fn,kn)\}$$

…… Eqn (2)

The f1, f2, f3…fn are the files and k1, k2, k3…kn are the keywords used for the processing of the word location based indexing. The query result using WLI-SP model is fetched using the data list and data tables. The Merkle tree based search in WLI-SP model reduces the memory usage quickly and fetches the result very easily

Sequence Pattern Mining in WLI-SP Model:

Once the word location indexing is performed efficiently using the Merkle tree, the second process to be followed is the process of sequential pattern mining. Sequential pattern mining in word-location based indexing is used to handle both on the shorter and longer user queries in deep mining. Let 'Seq' be a sequence of 'n' files with spatial locations {loc0, loc1, . . . , locn−1} representing the movement of an queries over a long history in web database. A sequential segment 'seg' is defined by a subsequence $loc_i, loc_{i+1}, loc_{i+T-1}$ of 'Seq' such that 'i' modulo where T=0. T denotes the number of file 'f' terms used for similarity measurement.

The number of query transactions in a deep web mining is called the length of the sequence. A sequence with length 'l' contains β, and it is denoted as length (β) = l. Furthermore, 'i' itemset is denoted by β[i] and prefix deep word-location mining process is used to process several times using different set of itemsets in a sequence. The length of the sequence pattern mining specifies the length of the user query terms obtained for the transaction in deep web mining. Length constraints in WLI-SP model also specified the number of distinct items, or even the maximal number of items per transactions. The user query also combines 'and' and 'either-or' query form together, so that the length of the sequence pattern increases gradually.

Prefix deep word-location mining process mine and answer the user queries clearly with query form such as 'and', 'or', 'neither-nor', 'either-or'. The

prefix deep word-location mining algorithm effectively discovered patterns utilizing the depth-first traversal strategy to clearly answer the user queries. Prefix deep word-location mining process takes the pattern growth to discover the frequent items after scanning the series of web database once. To conclude, the absolute set of Prefix deep word-location sequential patterns is established in a recursive manner for growing subsequence fragments in each projected web database. The web database scanning using WLI-SP model is shown in below Fig.3

File ID	Keyword ID	Sequence ID	Sequence
1	1:2	10	<sTd(a,b,c),(a,c),(a,d,c)>
2	2:4	20	<Day(a,c),(a,b,c), (b,c)>
3	3:5	30	<Night(a,d), (a,b), (a,c)>
4	4:6	40	<dTs(c,d,c),(c,a),(c,b,a)
5	5:2	50	<Cost((a,b,c),(a,c),(a,d,c)>

Fig. 3: Web database scanning using WLI-SP model.

Fig 3 briefly explains the web database scanning using WLI-SP model. File ID, Keyword ID, Sequence ID are used in deep web mining to fetch the user query result easily. For example, the keyword ID 1:2, denotes with the file ID '1' and '2' is the place where the keyword is placed in user query. In a similar manner, all the places of keyword in the file are denoted. <File ID; Position of Keyword in user query> denotes the all keyword ID effectively. For instance, the sequence pattern denotes the flight ticket reservation from the source point to the destination through different intermediate points.

The 'sTd' denotes the source to destination, where the source point in sequence ID is 10 is and destination denotes the point 'c'. Sequence ID 20 denotes the day flights going through the particular source and destination points. Similarly, sequence ID 30 denotes the night flights, sequence ID 40 denotes destination to source flights and sequence ID 50 points the cost fare of different routes. Prefix deep word-location mining process uses the depth first search to easily mine the user query.

Word-Location based Indexing Using Sequential Pattern Algorithmic Steps:

Finally the WLI-SP model performs effective indexing based using the data list and tables. The algorithmic step involved in WLI-SP model is explained below:

Input: User Query 'UQ', URL, Title, Header {h1, h2, h3…hn}
Output: Effective indexing and answer user queries
Begin
//Word location indexing
Step 1: User Query for deep web mining
Step 2: Identification of Meta location using Merkle Search
2.1: URL: {Domain, Sub domain name}
2.2: Title: {Keyword in Query}
2.3: Header: {h1, h2, h3, h4….hn}
Step 3: Identify the similarity measure using Cosine similarity in word-based indexing between the Files 'F1', 'F2' ….'Fn'
//Word-location index using Sequential pattern
Step 4: Sequential of 'n' files with spatial locations
Step 5: Initialize the length of user query sequence as length (β) = l
Step 6: Discover the frequent items using the Prefix deep word-location mining process
Step 7: Apply Depth-first Strategy for clear query answer in deep mining
End

WLI-SP algorithm returns the user answer very clearly though the deep web mining. The preferences are organized and prefix deep word mined using the depth-first strategy. The depth first search in the Merkle search tree starts the search from the first word to the end of the word from the query to identify the location. The time and space is deeply analyzed in the experimental factors to fetch the best result in the deep web mining.

II. *Wli-Sp Experimental Setup:*

Word-Location based Indexing Using Sequential Pattern (WLI-SP) model is developed in JAVA using weka tool. The WLI-SP model performs experimental work which is based on the user queries. WLI-SP model takes the MSNBC.com Anonymous Web Dataset for experimental work. The MSNBC.com Anonymous Web Dataset contains the page visit count and also records visited in a specific URL. The URL category contains the domain and sub-domain link with 5.7 average number of visit per user and the search ranges from 10 to 5000.

Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com contains the series of information in the dataset which communicate throughout the 24*7 period. Each event in the sequence corresponds to a user's query request for a page to attain the clear result after the deep web mining. Requests are not recorded at the premium level of feature (i.e.,) at the level of URL are rather, recorded at the level of page category. The parametric factors taken for the evaluation are relative runtime, averaged accuracy ratio, deep mining query result ratio, true positive rate, memory usage and sequential pattern based similarity measure.

The relative runtime on Word-Location based Indexing Using Sequential Pattern is the amount of time taken for the user query response with clear result using deep web mining. WLI-SP model is compared with the other existing HyLiEn (Hybrid approach for automatic List discovery and Extraction on the Web), and Visually Extracting Data Records (VEDR) and measured in terms of seconds (sec). Memory usage factor as measured in WLI-SP model, describe as the amount of memory consumed to store the user query, measured in terms of Mega Bytes (MB). Sequential pattern based similarity is measured as the similarity and clear result obtained on the depth first search strategy.

Averaged accuracy parameter in WLI-SP model indicates the extent to which it is able to accurately relate to the user query information using web database information. WLI-SP model effectively relate the averaged accuracy and measured in terms of percentage against the existing system. Deep mining query result ratio defines the effective query terms which matches with the meta-location points. The meta-location analyzes the URL, title and headers. True positive rate factor measures the proportion of actual positives values which are correctly identified, measured in terms of success rate.

III. *Result Analysis Of Wli-Sp Model:*

Word-Location based Indexing Using Sequential Pattern (WLI-SP) is compared against the existing HyLiEn (Hybrid approach for automatic List discovery and Extraction on the Web), and Visually Extracting Data Records (VEDR). The value of table (1) evaluation given below and the graph describes the WLI-SP model that attains the clear result when compared with the existing system.

Table 1: Tabulation of Relative Runtime.

User Count	Relative Runtime (sec)		
	HyLiEn	VEDR method	WLI-SP model
25	333	285	241
50	455	443	353
75	545	474	402
100	603	499	428
125	741	626	551
150	849	740	628
175	947	843	748

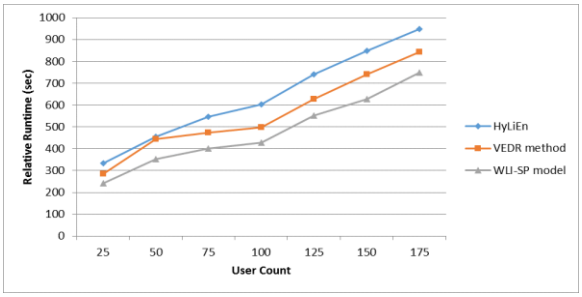

Fig. 4: Measure of Relative Runtime.

Table 1 figures out the relative runtime using the WLI-SP model, HyLiEn and VEDR method against the user count. The user count of 25, 50, 75 and 100 are taken for the experimental evaluation.

Fig 4 describes the relative runtime where the runtime percentage is reduced by 21 – 29 % in WLI-SP model when compared with the HyLiEn (Fabio Fumarola, 2011). The runtime relativity is decreased in WLI-SP model using the word location indexing that retrieve the relevant documents in deep web mining. Word location indexing uses the Merkle word location search tree, so that the user query result fetched very quickly. The easy fetching of result reduces the query runtime by 11- 20 % in WLI-SP model when compared with the VEDR method (Neil Anderson and Jun Hong, 2013).

The average accuracy ratio values of WLI-SP model, HyLiEn and VEDR method is shown in Table 2. As the iterations raises, the accuracy ratio is gradually improved.

Table 2: Tabulation for Averaged Accuracy Ratio.

Iterations	Averaged Accuracy Ratio (%)		
	HyLiEn	VEDR method	WLI-SP model
1	68	75	80
2	69	77	81
3	70	78	82
4	71	80	83
5	72	81	85
6	73	82	86
7	75	84	89

The averaged accuracy measure is experimented based on the iterations. The result fetched against different query term attains closeness in the result using the Merkle search tree with meta-location. The meta-location makes use of the URL, title and header. The cosine similarity with Merkle search tree improves the accuracy rate by 16 – 18 % when compared with the HyLiEn (Fabio Fumarola, 2011) and 3 – 6 % improved when compares with the VEDR method (Neil Anderson and Jun Hong, 2013). The results developed search results and computed through cosine similarity.

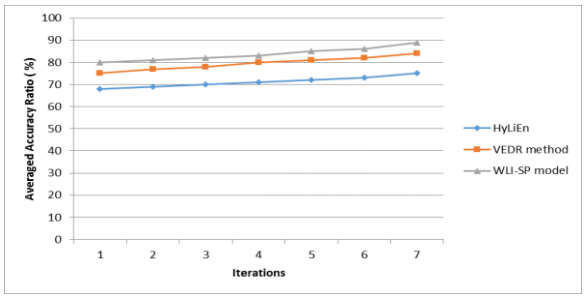

Fig. 5: Averaged Accuracy Ratio Measure.

Table 3: Tabulation for Deep Mining Query Result Ratio.

File Name with size (KB)	Deep Mining Query Result Ratio (Pass %)		
	HyLiEn	VEDR method	WLI-SP model
F1 (2)	60	72	90
F2(5)	61	73	91
F3(4)	62	74	92
F4(8)	62	75	93
F5(12)	62	77	93
F6(15)	63	78	94
F7(22)	65	79	95

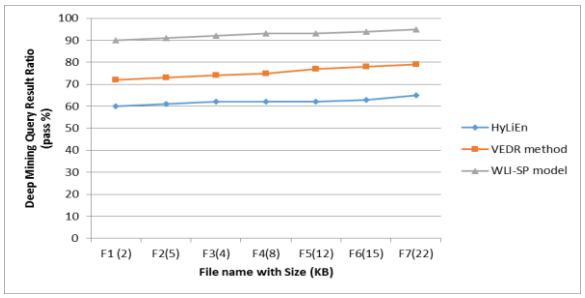

Fig. 6: Deep Mining Query Result Ratio Measure.

The query result ratio (as illustrated in table 3 and fig 6) is fetched using the different files name such as 'F1', 'F2' ….'Fn'. Each file size taken for the experimental evaluation is shown with file name and measured in terms of (KB). As the file size increases, the deep web mining query result is improved using the specified threshold value that easily selects the best result with meta-location. The best result with meta-location provides very easy result with 45 – 50 % improved query result ratio when compared with the HyLiEn (Fabio Fumarola, 2011) and 20 – 25 % improved query result when compared with the VEDR method (Neil Anderson and Jun Hong, 2013). . The word location indexing provides very easy result by selecting the keywords, URL and headers from the user query and performing the process.

Table 4 shows the true positive rate that is measured against the web users and positive rate measures the success rate. The success rate denotes the amount of clear result to the user query. As the user count increases, the true positive rate is also gradually increased.

Fig 7 describes true positive rate and rate improved by 3- 5 % when compared with the HyLiEn (Fabio Fumarola, 2011) and 9 – 13 % improved when compared with the VEDR method (Neil Anderson and Jun Hong, 2013). Prefix deep word-location mining algorithm effectively discovered patterns and utilize the depth-first traversal strategy to clearly answer the user queries with higher success rate. Prefix deep word-location mining process considers the pattern growth to discover the frequent items and improve the positive

rate when compared to the VEDR method (Neil Anderson and Jun Hong, 2013).

Table 5 and Fig 8 illustrates the memory usage based on the query size. The query Size is measured in terms of Mega Bytes (MB). The tree based search in WLI-SP model reduces the memory usage quickly and fetches the result very easily. The query result using WLI-SP model is fetched using the data list and data tables. The space is deeply analyzed in the experimental factors and fetches the best result with 17 – 29 % improved when compared with the HyLiEn (Fabio Fumarola, 2011) and 8 – 15 % improved in VEDR method (Neil Anderson and Jun Hong, 2013) in the deep web mining

Table 4: Tabulation of True Positive Rate.

No. of users	True Positive Rate (Success Rate)		
	HyLiEn	VEDR method	WLI-SP model
20	88	83	91
40	90	84	93
60	91	85	95
80	92	85	96
100	93	86	96
120	93	86	97
140	94	88	99

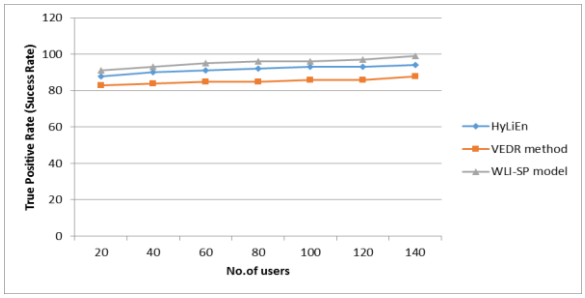

Fig. 7: Measure of True Positive Rate.

Table 5: Tabulation of Memory Usage.

Query Size (KB)	Memory Usage (MB)		
	HyLiEn	VEDR method	WLI-SP model
2	8.12	6.55	5.85
4	10.58	8.12	7.45
6	12.12	10.56	8.98
8	14.7	13.36	12.12
10	15.54	14.45	12.45
12	18.63	18.14	15.36
14	22.15	20.79	17.74

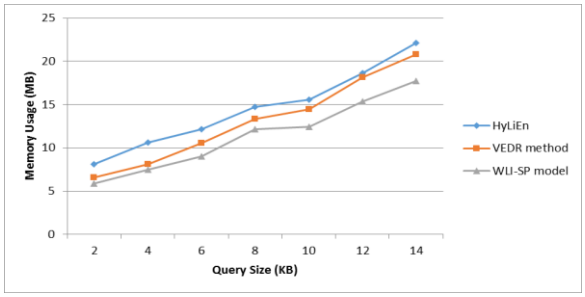

Fig. 8: Performance of Memory Usage.

Table 6: Tabulation of Sequential pattern based Similarity Measure.

No. of users	Sequential pattern based Similarity Measure (%)		
	HyLiEn	VEDR method	WLI-SP model
5	29	28	32
25	35	34	39
50	40	38	42
75	41	38	45
100	46	45	50
125	50	45	53
150	59	55	62

Table 6 and Fig 9 depict the similarity pattern based on the similarity measure in WLI-SP model. The cosine similarity is used individually for each file with the meta-location, so that the similarity between the files is identified easily. As the user count increases, similarity measure is improved by 5 – 12 % when compared with HyLiEn (Fabio Fumarola, 2011) and 10 – 18 % improved when compared with the VEDR method (Neil Anderson and Jun Hong, 2013). The depth first search in the Merkle search tree starts the search and improves the similarity score.

As final point, WLI-SP model retrieve the user result effectively with Merkle search using the cosine similarity measure. Sequential pattern growth in word-location based indexing efficiently and proficiently pushed deep search to get into the user query for the clear query result. The experiment is performed using different parametric factors and attains the effective indexing on the data list and tables.

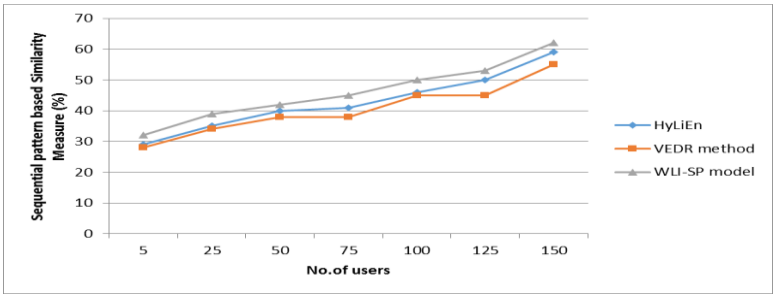

Fig. 9: Measure of Sequential pattern based Similarity Measure.

IV. *Related Work:*

The queries present in the Deep Web contents are processed and then provided to the web databases. The records that are returned are further enwrapped in web pages. Extracting structured data from deep web pages is the most concerned problem due to the fundamental intricate structures of such pages. A confront linked with deep web systems, which has not established attention so far, arises because the deep web databases within a definite domain are frequently not independent, i.e., the output results from one database are needed for querying another database. For a given user query, multiple databases may require to be queried in an efficient order to recover all the information desired by a user.

Thus, there is a need for techniques that generate query plans, accounting for dependencies between the data sources and extracting all information preferred by a user. The deep Web, which surfaces database records as web pages in (Richi Nayak, 2012), is not effective on structure and content features for appropriately mining the data using DSML. It is also challenging to combine structure and content features using different types of data models.

Deep iCrawl as presented in (Anita, R., 2011) splits the process into two user query analysis and user query translation for extracting data from the deep web. The second envelop vision-based extraction of data from the dynamically created deep web pages is not effective in integrating the features. Vision-based Page Segmentation (IVBPS) algorithm as demonstrated in [8, 9] expands intelligent crawler and interpretation of hidden web query interfaces. Hidden intelligent crawler and interpretation of hidden web query interfaces divide the method into two stages. The primary stage comprises user query analysis and user query translation and the subsequent wrap vision-based extraction of data from the dynamically formed hidden web pages.

Deep web data integrative configuration classify web database by domain, to present users with an integrated query interface, called the integrated interface. Web database query interface itself, is called the local interface. Through the query interface, users present queries to many local interfaces of web databases at the same time. Mapping user queries with homogeneous interface to the local interface is the key concern in pattern matching of components. Visually extracting records from deep web as demonstrated in (Neil Anderson and Jun Hong, 2013) uses the common sources of evidence to recognize data records on a displayed query result page. Based on these web observations, it recognizes every data record separately such as navigation bars but data items are not semantically labeled in tables.

V. *Conclusion:*

Word-Location based Indexing Using Sequential Pattern retrieved clear user result on deep web mining. The WLI-SP model used the meta-location with Merkle search tree. The Merkle search with cosine similarity finds the similarity between the files used for user query analyzing. The cosine similarity is used individually for each file with the meta-location, so that the similarity between the files is identified easily. Sequence pattern using the prefix deep word-location mining process uses the depth

first strategy. The depth first strategy in the prefix deep word-location mining process analyzes the time and space factors effectively. Prefix deep word-location mining process takes the pattern growth to discover the frequent items and answer user query. The experimental study results obtained by WLI-SP model attain the minimal runtime, and memory usage. The WLI-SP model on JAVA Weka tool achieves 5.215 % higher accuracy ratio, deep mining query result ratio, true positive rate, and sequential pattern based similarity measure.

REFERENCES

Fabio Fumarola, Tim Weninger, Rick Barber, Donato Malerba and Jiawei Han, 2011, "Extracting General Lists from Web Documents: A Hybrid Approach," Springer-Verlag Berlin Heidelberg.

Neil Anderson and Jun Hong, 2013. "Visually Extracting Data Records from the Deep Web," ACM journal.

Trupti, B., Mane and Prof. Girish P. Potdar, 2012. "Template Extraction from Heterogeneous Web Pages," International Journal of Advanced Computer Research, (ISSN (print): 2249-7277 ISSN (online): 2277-7970), 2(4): 6.

Oswalt Manoj, S., Nisha Soms and N.V. Shibu, 2011, "Visual Architecture based Web Information Extraction," Bonfring International Journal of Data Mining, 1. Special Issue.

Hung-Leng Chen, Ming-Syan Chen and Su-Chen Lin, 2009. "Catching the Trend: A Framework for Clustering Concept-Drifting Categorical Data," IEEE Transactions on Knowledge and Data Engineering, 21(5).

Anita, R., V. Ganga Bharani, N. Nityanandam and Pradeep Kumar Sahoo, 2011. "Deep iCrawl: An Intelligent Vision-Based Deep Web Crawler," World Academy of Science, Engineering and Technology, 5.

Kenneth Wai-Ting Leung and Dik Lun Lee, 2010. "Deriving Concept-Based User Profiles from Search Engine Logs," IEEE Transactions on Knowledge and Data Engineering, 22(7).

Kopal Maheshwari, Namrata Tapaswi, 2013. "Advance Frameworks for Hidden Web Retrieval Using Innovative Vision-Based Page Segmentation," IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727, 12-3.

Raghu, D., V. Sridhar Reddy, Raja Jacob, Ch., 2011. "Dynamic Vision-Based Approach in Web Data Extraction," D. Raghu *et al*, / (IJCSIT) International Journal of Computer Science and Information Technologies, 2(6).

Sawan Bhawsar, Kshitij Pathak, Sourabh Mariya and Sunil Parihar, 2012. "Extraction of Business Rules from Web logs to Improve Web Usage Mining," International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459), 2-8.

Boutros R., El-Gamil and Werner Winiwarter, 2013. "Accelerating Structured Web Crawling without Losing Data," ACM journal.

Mr. Vinod Kumar Raavi and Satya P. Kumar Somayajula, 2012. "Automatic Template Extraction from Heterogeneous Web Pages," International Journal of Advanced Research in Computer Science and Software Engineering., ISSN: 2277 128X. 2-8.

Wei Liu, Xiaofeng Meng and Weiyi Meng, 2010. "ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE Transactions on Knowledge and Data Engineering, 22(3).

Daiyue Weng, Jun Hong and David A. Bell, 2012. "Web Data Extraction from Query Result Pages Based on Visual and Content Features," International Journal of Software Informatics, 6-3.

Roberto Panerai Velloso, Carina F. Dorneles, 2013. "Automatic Web Page Segmentation and Noise Removal for Structured Extraction Using Tag Path Sequences," Journal of Information and Data Management, 4(3).

Srikantaiah, K.C., M. Suraj, K.R. Venugopal and L.M. Patnaik, 2013. "Similarity based Dynamic Web Data Extraction and Integration System from Search Engine Result Pages for Web Content Mining.," ACEEE Int. J. on Information Technology, 3(1).

Leng Goh, Jer Lang Hong, Ee Xion Tan and Wei Wei Goh, 2013. "Region Based Data Extraction," Communications of the IBIMA, Article ID 743515, DOI: 10.5171/2013.743515.

Panagiotis Papadimitriou and Hector Garcia-Molina, 2011. "Data Leakage Detection," IEEE Transactions on Knowledge and Data Engineering, 23(1).

Weifeng, S.U., Bnu-Hkbu, Pku-Hkust, Jiying Wang and Frederick H. Lochvsky, 2009. "ODE: Ontology-assisted Data Extraction," ACM journal.

Richi Nayak, Pierre Senellart, Fabian M. Suchanek and Aparna S. Varde, 2012 "Discovering Interesting Information with Advances in Web Technology," SIGKDD Explorations. 14-2.