Web Pattern Mining using ECLAT

Poonam P. Doshi
Research Scholar in Computer Engg.
Pacific University, Udiapur

Emmanuel M., PhD
PICT, Sr. no 27, Pune-Satara road, behind BV College, Dhankawadi, Pune, Maharashtra 411043

ABSTRACT
The use of internet has been increasing day by day. The users can find their resources with the help of different hyperlinks. These usages of Internet have led to the invention of web crawlers. The search engine which helps the user to explore the web is known as Web Crawler. In web crawlers the crawled data can be used to find missing links, community detection in complex networks. The concept of providing accuracy for this is forever in the vein. In this paper, web crawlers: their architecture, process of semantic focused crawling technology, ontology learning, pattern matching, types and various challenges being faced when search engines use the web crawlers, have been reviewed.

The web results more relevant to the user query through keyword expansion have been retrieved by the system. This data is being use further for the efficient association rule mining using Eclat Algorithm which is weaved for the vertical transactions based scheme. This process is being powered with Shannon information gain to identify the important words for the frequent pattern mining, and the whole process is being catalyzed by the fuzzy logic classification for more mere pattern identification process.

General Terms
Semantic crawling, Pattern mining, Algorithms

Keywords
Web crawler, Shannon information gain, Association Rules, Eclat Algorithm and Fuzzy.

1. INTRODUCTION
There is enormous development in the use of internet through World Wide Web (WWW). It contains a huge amount of information with the consistent inclusion of new data. There is enormous development in the use of internet through World Wide Web (WWW). It contains a huge amount of information with the consistent inclusion of new data [1]. It is projected that there were about 3.2 billion web clients till 2016, with an estimate of a yearly development more than 20%. According to the user query, it is vital to categorize the data as important or unimportant. 8

The extraction of the relevant information plays an important role. The Crawler is a program which reads and visits web pages for traversing and indexing web information based on domain, application, and the query. There are different types of crawlers as:

Web crawler: Web crawler: A web crawler (called a web bug or web robot) is a program/robotized content which scrutinizes the World Wide Web in a systematic and mechanized way, which is called Web crawling. Authentic sites, in specific search engines, use crawling as a means of providing latest data.

Focus crawler: In the focused crawler, website pages are gathered through the web with a specific property, via deliberately organizing the crawl frontier and taking care of the hyperlink investigation process. Thus focused crawler saves the bandwidth required to download a web page. The focused crawlers allow the reduction of the data collection and integration efforts. It crawls a specific part of the web to retrieve the relevant source.

Semantic focus crawler: It is a crawler which can search and download the page automatically for the given keyword. With the help of Semantic Web Technology, focus crawler can traverse the web, retrieve information from the web and download relevant information from the web. The aim of the semantic focus crawler is to get an efficient result and exact information to the given query [9]. In addition, data can be robotically updated in the crawling process. The semantic centered crawler is a subset of a centered crawler.

With the help of crawling, data can be crawled and subsequent index can be performed to the fetched document. The key components of a web search engine are Crawler and an Index [10]. The search engine has following components:

1.User Interface: It is user interface where the user can put URL to view the query result. 2. Web Crawler: Crawling procedure includes visiting web pages, parsing web data in a user-friendly format and indexing it properly. Pages are traversed on the web, searching links, sub-links to build important keywords and index generated data with the help of indexing. The Crawler initially accepts in URL as the source of information and iteratively finds subsequent URL's on the page and vital information. An Output set of the crawler is tree structure of URL's with seed URLs at the base.
The major process of crawling typically consists of the recursive process of the core as shown in the figure below:

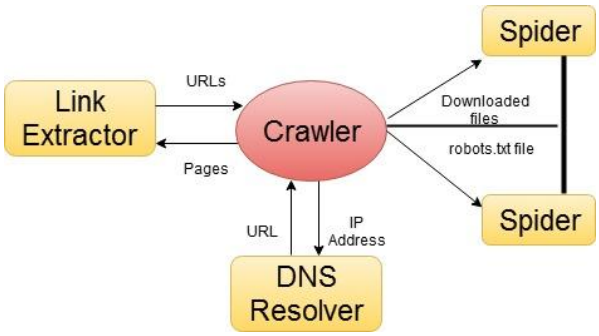

Figure 1: Web Crawler working

The performance and efficiency of web crawler depend not only on URL's set but also on different servers, their locations, data structure, the structure of repository and parsing technique [2]. Enhanced system performance has been

assisted by improved networking technologies. The system performance in the research is augmented by the key tasks of Parsing and Preprocessing. Parsing and preprocessing play the noteworthy role as it perks up the framework execution.

1. **Parser:** The Parser is the process which Scrutinizes web content by the riddance of HTML tags, JavaScript and redundant words from the information substance. The parser is the technique sorting information as natural text which contains nouns, pronouns, verbs and adverbs with an input rule set. Unprocessed information is altered into well-formatted information by the Parser.

The yield of Parser is specified by the Cleaning and Filtration process that engages the steps as follows:

- Deletion of high-frequency words is needed through the clear out with the eradication of the common stop words ("A", "am", "the", "is", etc.).
- Stemming is requisite for the morphological scrutiny to trim down the resultant words derived words to root words (fishing: fish).
- Tokenization technique is required to segment & break-up stream of text in atomic words, phrases or expression to generate meaningful elements termed as a token.

2. Data block content: Additional information should be sifted consolidating vital information just by observing likelihood of word to be found in web content. In decision tree process of learning information, the gain is proportional to characteristic data. Information gain is frequently employed to decide which of the features are most applicable so as to be practiced near the root of the graph.

3. Support value calculation: Shannon information equation facilitates calculation of entropy. Subsequent frequent itemsets are needed to be found in content which derives power set. In this process, a threshold is set and every terms weight is evaluated against a threshold value. A repetitive or inductive learning process is implemented to eliminate item sets from information content which are below ideal value. Powerset scanning each and every dataset item is generated to generate frequent item set.

Pattern mining leads to the discovery of the new patterns.

4. Pattern Match: New patterns are discovered in pattern mining, to find the associations and the relations among variables in the huge database. The association rules encourage in building valuable novel relations. Frequent itemset mining algorithms like Eclat facilitate support calculation and build confidence in web information. Information Retrieval system decision support is required to acquire optimal solution in available choices and state of condition.

Importance of Web Crawlers [6]
It is observed that 33% of the data is listed into the indexer. Firstly the downloaded data is the most relevant and excellent pages, which is the prime requirement of the client. If the web page is considered as the node having information between web pages and URL's; Data structure may be visualized as a graph. The crawler will require traversal mechanisms to navigate the graph. In the proposed crawler Depth First Search (DFS) approach is considered. This research paper has been organized in five sections Introduction, Related work, proposed Methodology, results in Discussion &Analysis& Conclusion.

2. RELATED WORK

A methodical writing survey and overviews are available in the literature to discover suitable calculation system and outlining the structure in order to solve the issues [2]. Literature is also available in web mining, pattern analysis, preprocessing, Éclat and apriori algorithm, web crawler, power set generation etc.

Today WWW contains millions of unstructured information useful for the users, many information searchers usage search engine to initiate their Web activity. Every search engine rely on a crawler module to provide the grist for its operation [18], Matthew Gray wrote the first Crawler, the World Wide Web Wanderer, which was used from 1993 to 1996 [19]. J. Cho. in [18] describes various search techniques and how the search engines works by using crawler and in [20] he has described how the search engines should cope with the evolving Web, in an attempt to provide users with up-to-date results.

Some of the issues are considered in Learnable Focused Crawling Approach. Pre-processing encourages better order and classification of information, this requires better information handling and thus pre-processing is the fundamental stage in web mining [7]. The researchers are trying to improve the performance of semantic focused crawlers. The objective of ontology learning is to extract facts or patterns semi-automatically from the quantity of data and transform these into machine-readable ontologies [8]. The diverse techniques designed for the ontology learning are statistics-based techniques, natural language processing based techniques, logic-based techniques and so on. These techniques are classified into semi-supervised techniques, supervised techniques, and unsupervised techniques commencing the viewpoint of learning control. The ontology-learning-based techniques can be used to solve the issue of semantic-focused crawling. The new information is learned from the crawled documents and the new information is integrated with the ontology. In order to calculate the 'relevance score', between topics and text document, an unverified ontology-learning based focused crawler [12] is used.

The frequent item in the huge data set is discovered and association rule mining is used. Numerous methodologies and techniques have been used to discover frequent item [13]. The subsequent table describes the patterns which could be mined with association rules and techniques as shown below:

Table 1: Association rule Mining Approaches and Techniques

Approach	Core Technique
Apriori	FP-Tree
AprioriTID	Eclat
DHP SPADE	FDM SPAM
GSP	Diffset
DIC	DSM-FI
PincerSearch	Prices
CARMA	CHARMP

Researchers describe the bottomless outline of web crawler [14] that bots recursively for mine Web Information through the approval of web page URL.

A significant contribution of the author is on static data set to improve the efficiency and performance.

3. PROPOSED METHODOLOGY

The proposed research relates to a search engine, a method of construction & working within the framework of the semantically based crawler for the dynamic administration data over the web. The internet is the largest marketplace in the world for retrieve, store data of various platforms. The internet promoting is exceptionally well known for various businesses, including the conventional mining administration industry, where mining administration ads are effective carriers of mining administration data [1]. The architecture of the system is as shown in Figure 2:

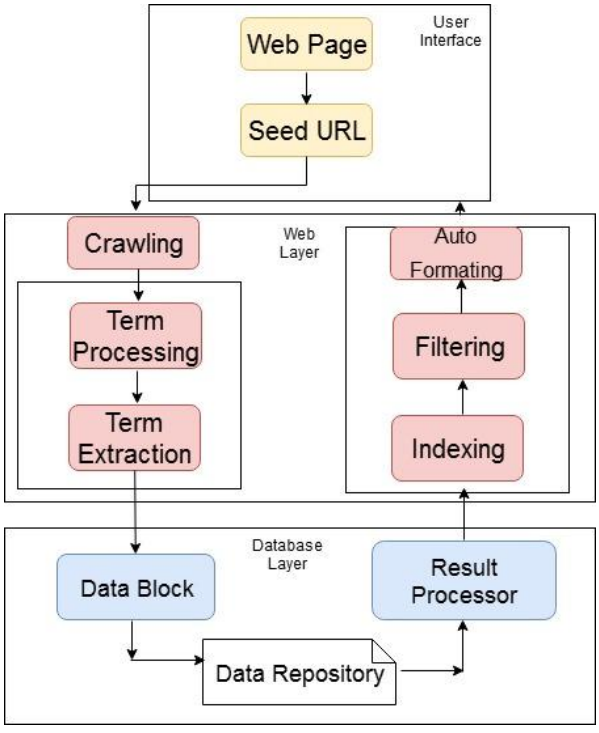

Figure 2: Proposed Architecture system [1]

Semantic focused crawling technology is utilized to resolve the matters of mining service information.

Let S={ } be a system for pattern recognition for web information.

A set S = {s1, s2, s3,….sn} where sn is seed URL.

To identify the interesting pattern S={sn, I}.

The system is divided into different stages and implemented with the help of algorithm to produce the pattern matching.

3.1 PROPOSED CRAWLING ALGORITHM

URLs from the frontline are recursively visited according to a set of policies. The crawling algorithm steps are as below:

Procedure1. :: Proposed Crawling algorithm

queue (Q) is initialize with initial set of known URL's.

Until Q empty or time limit drained:

Pop URL, U, from front of Q.

If U is not an HTML page (.gif,.jpeg,.ps, .pdf, .ppt)

exit loop.

If already visited U, continue loop (get next url).

Download page, I, for U.

If cannot download I

exit loop,

else.

Index I

Parse I to obtain list of new links N. Append N to the end of Q.

Procedure2. :: Depth First Crawlers

Input: graph G and a start vertex node of G

SetLabel(node, VISITED)

edge ∈2 G.incident Edges(node)

GetLabel(edge) = UNEXPLORED

weight = opposite(node; edge)

GetLabel(weight) = UNEXPLORED

SetLabel(edge, DISCOVERY)

DFS(G, weight)

SetLabel(edge, BACK)

Output: labeling of the edges of G in the connected component of node as discovery edges and back edges

Depth First Search (DFS) is a useful technique which traverses through the search by starting at the root link and traverse deeper through the child link. If there is more than one child, then priority is given to the leftmost child and traverse deep until no more child is available [1].

Crawled is essential as it retrieve raw information from desired links. This retrieved information is then subsequently passed to preprocessing required to clean and filter data which is crawled.

3.2 Preprocessing

Sentence Segmentation: Segmentation helps to recognize the boundary and generate set of sentences from information.

Tokenization: In tokenization, meaningful words are separated from above-separated sentences. Special symbol

removal algorithm will remove the special symbols like (@, #, $, %, &, ^, *, etc.)

High-frequency words: High-frequency words are quite simply those words which occur most frequently in a document, for example, "the ", "and ", "it ", “for”, “were”, “does”, "as ", etc. These are often words that have less meaning on their own, but they do contribute a great deal to the meaning of a sentence. This process eliminates the high-frequency words from the document, which reduces the size and do not affect document.

Stemming words: Stemming refers to the process of removing prefixes and suffixes from the words. In the preprocessing context, stemming is used to verify the word structure to avoid mismatches that may undermine the review. It is necessary to drive root words for meaningful information and reduction in time complexity (e.g. taking: taken: take) here suffixes like “ing”, “en” are eliminated.

Output: Processed meaningful Information ready to be cleaned and filtered.

3.3 i) String Matching

In String matching will be done and if the new string is found important then that will be sent to the updater.

ii) The Updater

Updater will be updated database repository for new strings found.

iii) Data Repository

This is the actual database in which all data will be stored.

iv) Result Processing

Result processing module will take the result from the database repository for processing and sending it further for integration.

3.4 Power Set Generation

After preprocessing, significant words are deposited in the data repository. The string matching with the help of ontology learning is done and the new string is found, which is sent to the updater. Indexing method is applied to have an index for the significant words, to facilitate fast and accurate information retrieval. Web Indexing is a substitute name for the procedure with the context to web indexes and designed to discover site pages on the internet.

The purposes of storing an index are to optimize the speed and the performance to find the relevant documents for a search query. To select most important data, data filtering and cleaning is done by using information gain is presented as:

$$I(s) = \sum_{i=1}^{n} \text{Probi} \log 2 \text{Probi} \quad \ldots\ldots \quad 1$$

3.5 Eclat algorithm

The output of frequent itemset is given further to Eclat algorithm for association mining. In each recursive call, the function intersects Tidsets, which verifies each itemset with tidset pair {P, t(P)} with all the others pairs {Q, t(Q)} to generate new candidates NPQ. If the new candidate is frequent, it is added to the set FP. To find frequent sets the algorithm searches in a DSF manner.

Recursive Learning concept is implemented and items having lesser confidence and support are eliminated for given threshold with Eclat. The procedure is as mentioned below:

Procedure 4. :: Eclat algorithm

Input: Alphabet M with ordering less than equal to multiset T ⊆ P(M) of sets of Itemm, Minimum support value minsup∈ L.

Step 1: Get database scanned for each item (Tidset)

Step 2: Tidlist of {x} is exactly the list of transactions containing {x}

Step 3: Intersect tidlist of {x} with the tidlists of all other

items, resulting in tidlists of {x,b}, {x,c}, {x,d}, …

= {x}-conditional database (if {x} removed)

Step 4: Repeat from 1 on {x}-conditional database

Step 5: Repeat for all other item m

Here Fuzzy logic is been implemented for set of five rules implicating on frequent itemsets. fuzzy logic is be functioned to mine precise frequent itemsets. input set of rules produced from Eclat´ procedure consists of frequent itemsets and support values.

4. USER INTERFACE SCREEN SHOTS

The following screen shot shows the crawling of the data successfully:

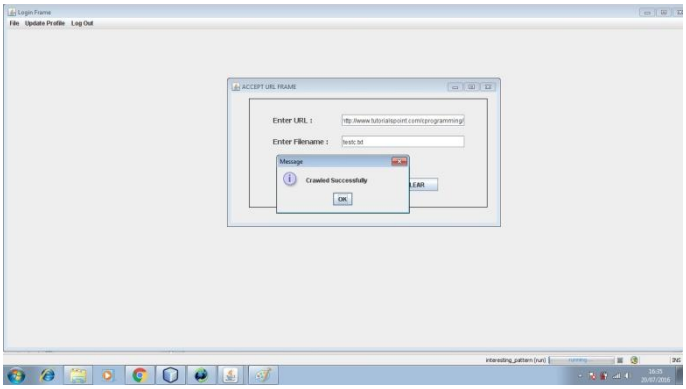

Figure 3: Crawling of the data successfully

The following screen shot shows the browsing of the crawled data.

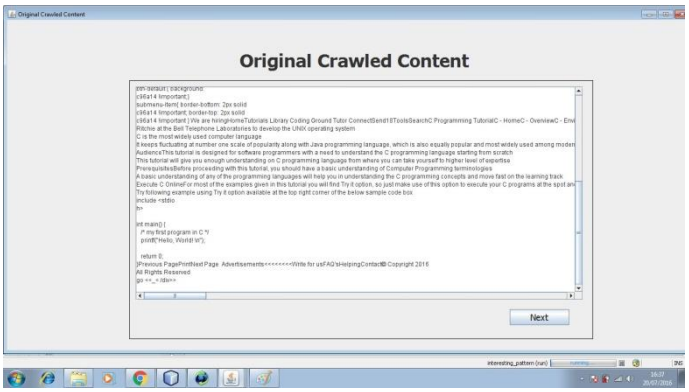

Figure 4: The original crawled data

The following screen shot shows the frequent pattern.

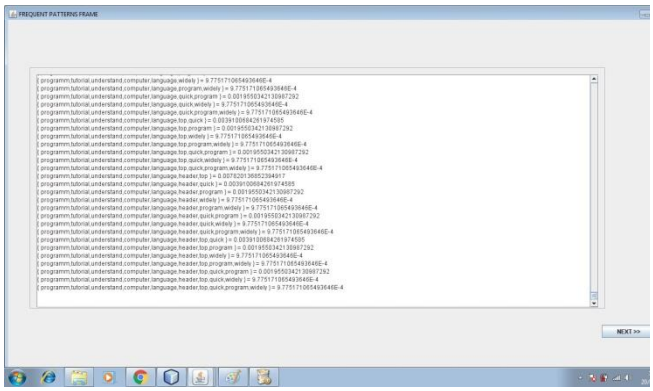

Figure 5: The Frequent Pattern

The following screen shot shows the Eclat pattern data.

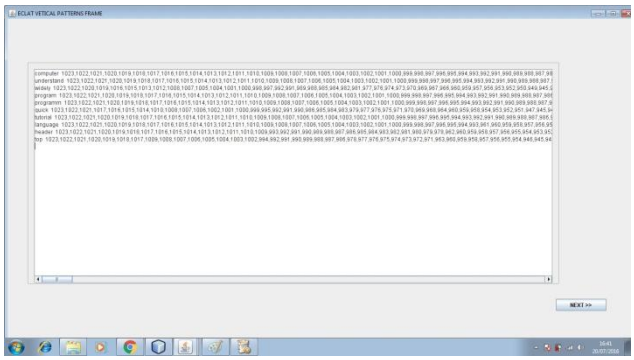

Figure 6: The Eclat Pattern data

5. TEST DATA AND RESULTS

System is being implemented in java on windows platform with machine configuration: 100HD 2GB Ram and tested for live web pages set. Interesting patterns found are recognized with fuzzy logic and Eclat algorithm. System is being evaluated against precision recall graph for performance.

Precision

Number of interesting patterns recognized to total number of relevant or irrelevant existing patterns. Precision give effectiveness of system in terms of percentile.

Recall

Number of relevant frequent patterns to sum of relevant patterns not recognized. abso-lute system performance is calculated by recall.

Consider following equation for computation:

P = The number of relevant Frequent patterns identified,

Q = The number of relevant Frequent patterns not identified,

and R = The number of irrelevant Frequent patterns identified.

So, Precision = (P/ (P+ R))*100

And Recall = (P/ (P + Q))*100

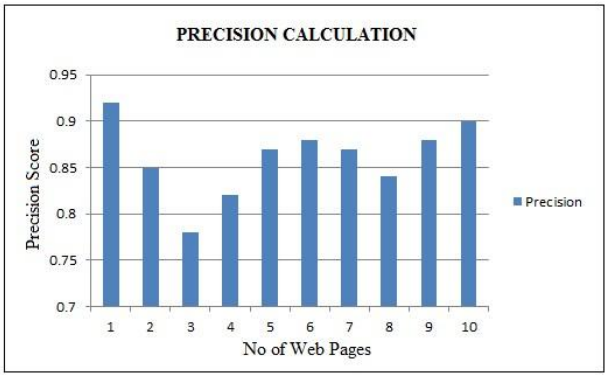

Figure 7: Precision

Table 2 : Average precision of the proposed approach

Evaluation	No of pages	Precision
I	1	0.925
II	6	0.86
III	10	0.9
Average		0.861

The average precision observed here is 3 evaluations is 0.85 which evaluates that system recognizes better interesting patterns.

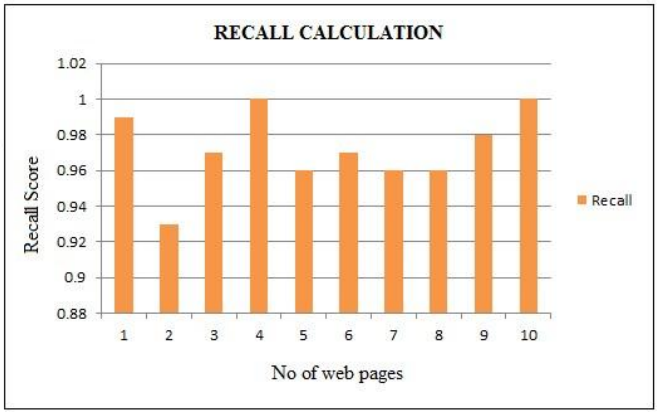

Figure 8: Recall

As found in above graph the average recall of system is 0.96 evaluated for set of 3 values.

Table 3: Average recall of the proposed approach

Evaluation	No of pages	Precision
I	1	0.99
II	6	0.95
III	10	1
Average		0.96

6. CONCLUSION

System sufficiently shows the better precision for the extraction of interesting patterns form the web pages. Efficiently extracts the web pages textual data and parses them to get rid of the redundant data. Then the parsed data is been preprocessed to get the most important data using Shannon information gain theory. System successfully identifies the all the possible frequent item sets with their candidates sets and this horizontal data is been converted into vertical data for Eclat´ mining algorithm. The results of the Eclat´ is been classified using Efficient rules of Fuzzy logic to identify Interesting patterns of the extracted web data for the effectiveness.

7. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the idea.

8. REFERENCES

[1] Poonam P. Doshi, Dr. Emmanuel M.: "Feature Extraction Techniques Using Semantic-Based Crawler for Search Engine", in Proc. of an International Conference on computing, communication and energy systems. ICCCES – 2016, in association with IET, UK and sponsored by TEQIP _ II, 29th 30th Jan 2016.

[2] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik "Study of Web Crawler and its Different types", OSR Journal of Computer Engineering ISSN 2278-8727 Volume 16 Issue 1 Feb 2014.

[3] H. Dong, F. Hussain, and E. Chang, O. Gervasi, D. Taniar, B. Murgante, A. Lagana, Y. Mun, and M. Gavrilova, Eds., "State of the art in semantic focused crawlers", in Proc. ICCSA 2009, Berlin, Germany, Vol. 5593, pp. 910–924, 2009.

[4] M. Ehrig, A. Maedche, "Ontology-focused crawling of web documents", in SAC'03: Proceedings of the 2003 ACM symposium on Applied computing, ACM Press, New York, NY, USA, pp. 1174–1178, 2003.

[5] A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic, R. Volz, "Ontology-focused crawling of documents and relational metadata", in Proceedings of the11th International World Wide Web Conference WWW-2002, Hawaii, 2002.

[6] Prashant Dahiwale, M. M. Raghuwanshi, Latesh Malik, "Design and Implementation of Focused Web Crawler Using Genetic Algorithm: An Approach to Web Mining", International Journal of Scientific & Engineering Research, Vol. 6, no. 6, June-2015

[7] Vijayashri Losarwar, Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012.

[8] W. Wong, W. Liu, and M. Bennamoun, "Ontology learning from text: A look back and into the future", ACM Computer. Surveys, Vol. 44, pp.20:1–36, 2012.

[9] Dong, H., Hussain, F. K., Chang, "E.: State of the art in semantic focused crawlers Computational Science and Its Applications", – ICCSA 2009. Springer-Verlag, Seoul, Korea (July 2009) pp. 910-924

[10] Soumen Chakrabarti, "Mining the Web Discovering knowledge from hypertext data", Boston: Elsevier, 2012.

[11] H. T. Zheng, B. Y. Kang, and H. G. Kim, "An ontology-based approach to learnable focused crawling", Inf. Sciences, Vol. 178, pp.4512–4522, 2008.

[12] C. Su, Y. Gao, J. Yang, and B. Luo, "An efficient adaptive focused crawler based on ontology learning", in Proc. 5th International Conference, Hybrid Intell. System. (HIS '05), Rio de Janeiro, Brazil, 2005, pp. 73–78.

[13] Slimani, Thabet, and Amor Lazzez. "Efficient Analysis of Pattern and Association Rule Mining Approaches", Ar Xiv preprint ar Xiv: 1402. 2892 (2014).

[14] Khurana, Dhiraj, and Satish Kumar. "Web Crawler: A Review", IJCSMS International Journal of Computer Science & Management Studies 12.01 (2012).

[15] Jain, Nidhi, and Paramjeet Rawat. "A Study of Focused Web Crawlers for Semantic Web", International Journal of Computer Science and Information Technologies 4.2 (2013): 398-402

[16] Sonali Abhane, P. D. Lambhate "Enriching Web Interesting Pattern Mining Using Vertical Transaction Process", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, 2016

[17] Dr. Emmanuel M., Mr. Saurabh M Khatri, Dr. Ramesh Babu D. R. "A Novel scheme for Term weighting in Text Categorization: Positive Impact factor", IEEE International Conference on Systems, Man, and Cybernetics, 2013

[18] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. "Searching the Web". ACM Transactions on Internet Technology, 1(1), 2001

[19] Dr Rajender Nath, Khyati Chopra, "Web Crawlers: Taxonomy, Issues & Challenges", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013.

[20] Ntoulas, A., Cho, J., and Olston, C. "What's new on the Web? The evolution of the Web from a search engine perspective",WWW04,1-12,2004.