International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Enriching Web Interesting Pattern Mining Using Vertical Transaction Process

Sonali Abhane[1], P. D. Lambhate[2]

[1]M.E (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India.
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

[2]Professor (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India.
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

Abstract: ***Identifying Interesting patterns from the web is especially a tedious task, as the web pages are having different narration styles and data is distributed vastly. Most of the systems are existed to conduct the experiments on this but the concept of providing accuracy for this is always in the vein. So as an initial step towards this proposed system uses an interactive web crawler to crawl the textual data from the web pages. This data is being use further for the efficient association rule mining using Éclat 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.***

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

1. Introduction

In Current Scenario Web is source of vital information. Almost 4.05 billion pages been indexed by search engines and total count of indexed pages remains non- estimated from 2006[3]. Searching and Retrieving Relevant information is major objective of computer Technology. Finding new patterns from web and retrieving information related to topic is major challenge for web mining domain.

Commercial search engines like Google, Bing have proved to be best innovation for Information search on web and process almost 2.5 pet bytes of data daily [3]. Keyword search has been reason of popularity and major technique employed by these systems. Expansions of web bring in new undiscovered patterns and many search applications cannot retrieve information from hidden web pages, fail to find diverse patterns. This urges for new better technique in web crawling.

Design and Development of better Web crawling system for finding new innovative patterns from web is research challenge.

A Better and Effective retrieval of information and finding interesting patterns from web, web crawlers are needed. Crawlers assist in retrieving information related to topic and make process of indexing better [1, 2].Crawler is software program which assists in indexing of web pages. Crawler and according to technique involved in operation there are four types of crawling process i.e types of crawler.[1,2]

[1] Focused crawling
[2] Distributed crawling
[3] Incremental crawling
[4] Parallel crawling

Crawling process involves visiting web pages, parsing web information in human readable format and indexing it appropriately. It traverses pages on web, finding links sub links to build a list of words and associated information index generation. Crawler initially accepts in URL as source of information and iteratively finds subsequent URL's in page and vital information. Output set of crawler is tree structure of URL's with seed URL's at base.

Major Process of crawling typically consists of below recursive process at core as shown in fig

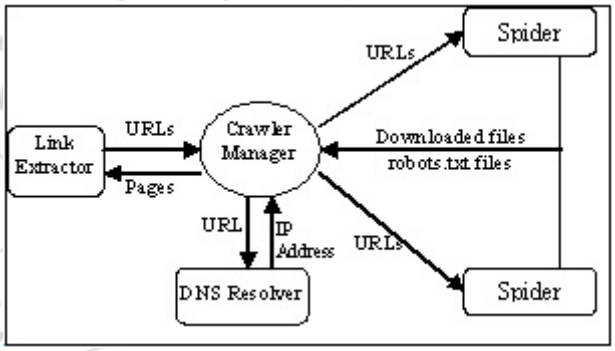


Figure 1: Web Crawler working [4]

[1] Download page
[2] Parse web page
[3] Repeat process for every seed URL

Performance and efficiency of web crawler depends not on URL's set but also different servers their locations, data structure, structure of repository and parsing technique. [2] Better networking technologies and large server numbers have already facilitated in better system performance. In this research scenario parsing and preprocessing remains major task in enriching system performance.

Parser analyzes web page content eliminating html tags JavaScript and bad words from information content. Parsing is technique classifying information as natural text consisting of noun pronouns verbs ad-verbs with input rule set. Parser

converts raw information to well formatted information. The output of parser is been given to cleaning and filtration process termed as preprocessing and involves following steps

[1] Stop word removal.
[2] Stemming.
[3] Tokenization.

Stop word removing is process of cleaning that eliminates common stop word like "the", "is" from information content. Many of word in information content are adjectives having superlative tone and fall short to make clear information meaning according to context. Stemming facilitates morphological analysis, reducing derived words to root words: fishing: fish: fish, advantages of stemming are query expansion.

Tokenization is technique to segment and break up stream of text in atomic words, phrases or expression and generating meaningful elements termed as token.

Further information needs to be filtered incorporating important data simply by finding probability of word to be found in web content. In decision tree process of learning information gain is ratio to intrinsic information which reduces biased meaning of attributes by counting number and size of branch in decision tree. Information gain relation biases decision graph against considering features with big number of distinct values. Solving downside of information gain namely information gain applied to features that could take on large numeral of separate values may discover training set too well. Information gain is frequently employed to decide which of features are most applicable so they could be practiced near root of graph.

For instance One of input features might be user's credit card number. This Features has a high information gain because it exclusively recognizes every user but deciding how to treat a user based on their credit card is improbable to abridge to customers which have not been used.

Shannon information equation facilitates calculation of entropy i.e gain in Information retrieval. Subsequent frequent item sets are needed to be found in content which derive power set. In this process a threshold is been set and every terms weight is been evaluated against threshold value. A repetitive or inductive learning process is been implemented to eliminate item sets from information content which are below ideal value. This generates powerset scanning each and every dataset item to generate frequent item set.

In pattern mining discovering new patterns we require to find associations and relations among variables in huge database. Finding interesting association among terms. Association rules facilitate in building useful novel relations. Frequent item set mining algorithms like Eclat facilitate support calculation and build confidence on web information that has been gathered. Today in every Information Retrieval system decision support is been required to acquire optimal solution in available choices and state of condition. Fuzzy logic facilitates simple IF-Else Fuzzy set to find solution set. All available input values are fuzzified in to fuzzy member functions and operated on all rules, lastly de-fuzzified to get crisp answers. Fuzzy logic has huge application is development of optimal system. This research paper has been organized in five section Introduction, literature survey proposed Methodology results discussion and conclusion

2. Literature Survey

A systematic literature review has been on IEEE, ACM International Journal and online web portals traversing through each and every selected ten articles finding issues and challenges . outcome of literature survey urges research work to find appropriate algorithm methodology and design framework to solve issues.

A. Survey Methodology

Survey methodology has been finding recent articles on every on every process related to web mining and pattern analysis.
Survey has been done on web mining and pattern analysis with Study and review of articles on "pre-processing" "éclat and apriori algorithm" "web crawler" "power set generation".

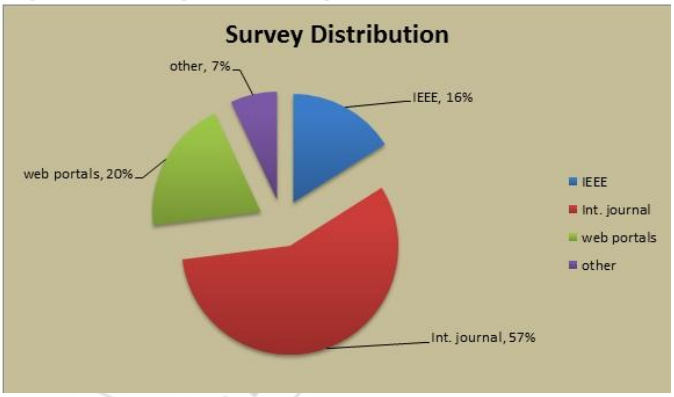


Figure 3: Survey Distribution

B. Survey Analysis

[5] Information present on web is mostly unstructured, dynamic, heterogeneous and diverse in pattern making tough for user to retrieve faster clean information additionally giving rise to scalability issue. Better pre-processing has been achieved with four step data fusion data cleaning and filtration module and found to best in survey on pre-processing techniques by author.

[6] Data on web has two distinctive features as to database. Exponentially high distributed and dynamic. Pre-processing facilitates better classification and categorization of information. This requires better data processing and hence pr-processing is vital phase in web mining.

[7] Common issue faced by éclat and Apriori is "field association". author has presented fuzzy logic to modify and solve this issue in case of apriori procedure. Huge dataset and outlining threshold to reduce perfectly result set has been major issue in apriori procedure. Fuzzy procedure

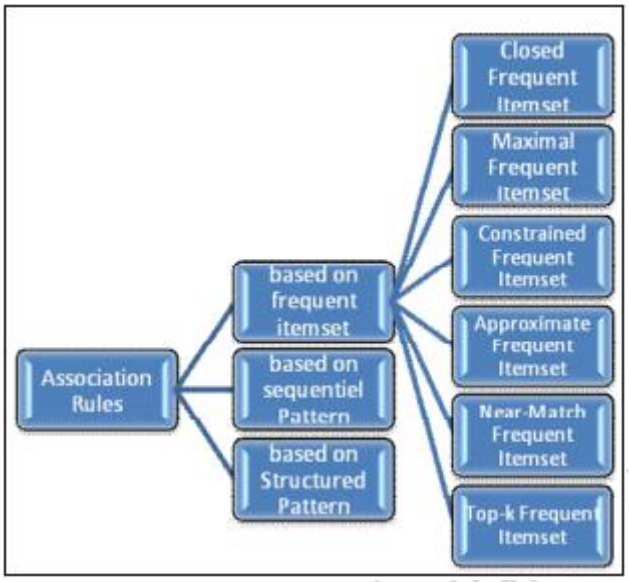


Figure 2: Association Rules [8]

Distributes data in various clusters and subsequently presenting user to give value for threshold. This technique considers each and every rule which is useful for interesting pattern mining.[9] Author presents enhanced algorithm with fuzzy based procedure for data mining that enhances accurateness and effectiveness of web mining process. algorithm employees Borgelt's prefix trees to generate association rule and is core technique in system. Confidence support and similarity has been computed with improved method approving efficiency of procedure. Generally, system is planned to recognize anomalous data usually persuaded between auditing network information.

As argued in above section discovering frequent item in huge data set association rule mining is been used. Numerous methodologies and techniques have been used to discover frequent item. [8] Analyzes various methodologies used for same. Subsequent figure describes patterns which could be mined with association rules. Article presents each and every pattern mining approach as below.

Table 1: Association rule Mining Approaches and Techniques

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

[11] Describes bottomless outline of web crawler. Web crawlers are bots that recursively for mine web information through acceptance of web page URL.

Article provides historic analysis of crawler i.e. In what way crawler evolved? Diverse data structures incorporated, different crawling techniques and in accordance different crawlers. Best article to understand concept of crawling issues and challenges related term. Major contribution of author has been detailed architecture view of different crawlers.

[10] Slug or Scutter, semantic crawler for web mining has been presented. Implementation of bot has been done with API in java (JENA).Slug build a framework for effective and consistent retrieval of web information from web pages. Vocabulary is been facilitated by program for enlarging web mining in better way. Metadata is been stored in crawling procedure for better web mining. Above information is been used by bot for observation and retrieval of better information. Figure 3 demonstrates working prototype of slug architecture.

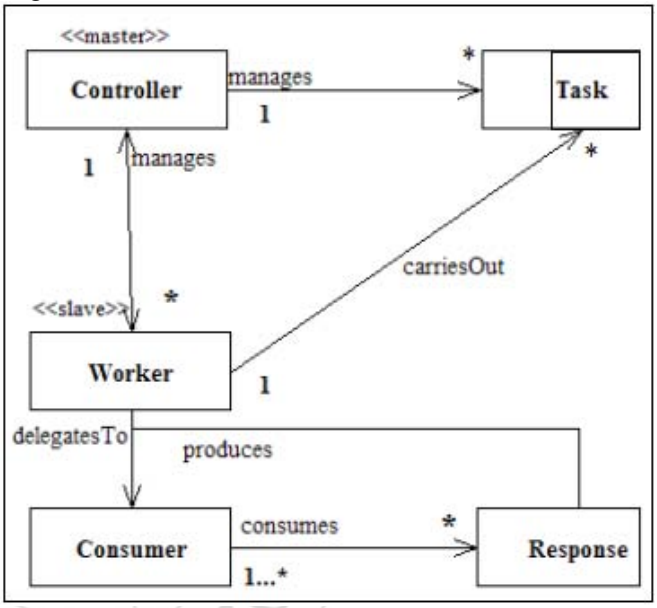


Figure 3: Slug Architecture [10]

Today with web 3.0 and further advancement in technologies For effective retrieval and store of information "ontology" technique has been widely accepted. Ontology maintain descriptive knowledge of entities and their association for every domain of categorization. [10] presents web crawler on ontology, here information of downloaded web page and ontology helps crawler to find desired location easily. Ontology makes retrieval and processing of data in better way and has found to be best in all overall techniques. It is the best and effective technique for effective information retrieval.

[12] Author has presented a profound study on semantic web and focused web crawlers. In process of crawling it is very challenging to make decision concerning to crawl a particular URL's to find pertinent data. Author has presented diverse readings and proposed techniques in last decade and each and every technique has been elaborated by author. Major contribution of author has been finding limitations and scope of work to overcome this limitation.

C. Issues and challenges: Finding Answers

This section summarizes review points found in survey to find issues challenges in research work and presents solution and techniques which can make enhance research

[5,6] Require Better Pre-processing: **4 phase pre-processing.**

[7,9] Association rule mining is challenge: **Apriori and Eclat**

[10,11] effective implementation of crawlers: **selection of correct crawler technique**.

[12] Web information distributed and diverse: **ontology**

D. problem definition

Keep it simple (KIS) principle has been used for defining problem definition: mining interesting web pattern from web.

3. Proposed Methodology

A. Core Methodology

Subsequent paragraph describes core methodology and architecture of system used for web information mining.

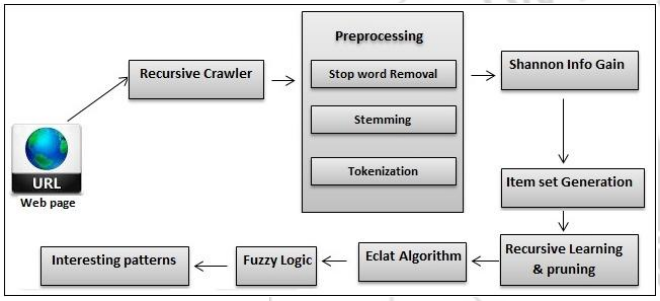


Figure 4: Proposed Architecture of system

The core methodology is five phase procedure and implements four major algorithmic procedure and subsequent phase.

Phase I: This phase system accepts set of seed URL related to particular domain from user. Each and every urls is been crawled and information is been stored in plain text files for future processing.

As information is been scattered in various web pages and stored on different types of servers. Gathering information related to user query pattern from web is challenging task, as merely mining information is not objective to achieve, but etrieval of interested pattern and relevant information is major goal to achieve.

Software bots i.e. programs like crawlers assist is retrieving information from distributed source and builds a structured tree i.e. graph based search for information to retrieve.

Proposed Web crawler implements recursive crawling technique, where in set of baby crawlers assist main crawler to retrieve subsequent urls and information stored on those URL in faster manner using multithreading.

Crawling process consists of:

[1] Download web content

[2] Parse content

[3] Outgoing links are parsed.

The above process is been repeated till program doesn't find any link for crawling and mining. Overall processing of crawler is graph based with seed link at base and other links as child nodes. Proposed crawler implements DFS algorithm for information retrieval.

Procedure [1]:: DFS (G, V)
Input: Graph "G" and Vertex "V"
Process:
SetLabel (v, VISITED)
For all e ∈ G.incidentEdges(v)
If getLabel (e) = UNEXPLORED
W ← opposite (v,e)
If getLabel (W) = UNEXPLORED
SetLabel (e, DISCOVERY)
DFS(G, W)
Else
Set Label (e, BACK)
Output:
Labeling of edges of "G" in connected component of "V" as discovery edges and back edges.

Crawler has been designed and developed on multithreading concept with java language. where many baby crawlers are replicas of main bot preforming repeitatative operation of visiting links and retrieving information. As this process involves complex and requires huge memory, a better data structure like vectors is been incorporated in implementation. Phase I is most vital as it assists in retrieving raw information from desired links of information's. In experimental scenario system has been tested for social networking portal. Main crawler retrieves information from web pages which has diverse format of English language and consists of JavaScript code and HTML tags. baby crawlers eliminate this unwanted information from raw information converting it readable format. advertisement and other unwanted things are omitted from retrieved information. This retrieved information is then subsequently passed to next phase II for preprocessing.

Phase II: here Information meaning is being interpreted word to phrases with 4 sub processes:

[1] Sentence segmentation
[2] Tokenization
[3] Stop word
[4] Stemming.

Procedure 2 is been used for preprocessing.

Procedure [2]:: Preprocessing

Input: Information from crawler

Process:
Start
Read string
divide string into records on space and store in a vector V
Remove Special Symbols
Identify Stop words
Remove Stop words
Identify Stemming Substring
Replace Substring to desire String
Concatenate Strings
stop
Output:
Processed meaningful Information ready to be cleaned and filtered.

- *Segmentation* helps to recognize boundary and generate set of sentences from information.
- *Tokenization* process separates meaningful words from above separated sentences.
- *Stop word:* conjunction play very less role in interpretation of sentence and very less usefull. This process eliminates words like ("is" , "the" , "for") and reduce extra complexity of extra processing.
- *Stemming:* In retrieved information numerous keywords are elongated and fail to convey meaning for given context. hence it is necessary to derive root words for meaningful interpretation of information and reduction in time complexity.(taking : taken: take) here ending suffix like "ing" , "en" are eliminated.

Phase III:
Data filtration and cleaning is been by selecting most vital information data with method of information gain as presented in equation (1).

$$I(s) = \sum_{i=1}^{n} - p_i \log_2 p_i \quad \text{.....(1)}$$

Here, p_i = evaluated probability of words weights for web pages.

Phase IV: Frequent items are generated from above filtered information generated with power set .Procedure 3 is been used power set generation.

Procedure [3]:: Power set generation

Input: Termset K_d and support threshold S0
Process:
Scaning Database F={Ø}
for(i=1; i< 2^ I; i++)
T_{count} =0;
for(j=1; j<2^I; j++)
If Ai ≤ Aj
$T_{count} = T_{count} + S_{count}$ (j);
If($T_{count} \geq$ S0)
Then F = F U Ai
Goto step 2
End
Output: list of frequent itemset.

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

Procedure [4]:: Eclat

Input: Alphabet S with ordering ≤ multiset T⊆P(S) of sets of Items , Minimum support value minsup Є ℕ.
Process:
G:={(Ø,|P|) }.
CØ:= {(x,P({x}))| x Є S}.
C'Ø:= freq (C ø):= {(x,P_x)(x,P_x) Є C ø, |P_x|≥ minsup }
G:= { Ø }.
Add frequent supersets (Ø, C'ø).
function add Frequent Supersets():
Output: Set G of frequent Itemsets and their support counts.

Input: frequent Itemsets fp Є P(S) called prefix, incidence matrix C of frequent 1-item-extentions of fp.

Process:
for (x, P_x) Є C **do**
q:= fp U {X}.
C_q:={(y,P_x ∩ P_v) | (y,P_v) Є C, y > x}.
C'q := freq(C_q) := {(y,T_v) | (y, P_v) Є C_q, |P_v| ≥ minsup }
If C'$_q$ ≠ Ø **then**
Add frequent supersets (q,C'$_q$).
End if
G := G U {(q, |P_x|)}
End for
Output: add all frequent extensions of fp to global variable G.

Phase VI:

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

Set will entail of diverse support values for items and itemsets are organized in ascending directive. 1 lowest support value than 1 highest support value from set will be

detailed. noted value will be splitted in 5 of range membership standards as:
[1] very low
[2] Low
[3] Medium
[4] High
[5] Very High.
Range is being ordered in low to high with IF-ELSE set of rules. Item set with interesting patters are recorded from this item set.

B. Mathematical Underpinning
Mathematical evaluation of system is done as below

I. PR= { } be as system for Patter recognition for web information
II. Identify Input as PR={ $U_1, U_2, U_3 \ldots U_n$}
III. Where U_n= Seed URL
IV. Recognize I as Output i.e. Interesting Patterns
V. PR= {U_n, I}
VI. Recongination Process P
VII. PR= {U_n, I, P}
VIII. P= {U_n, I, MW_c, PP_r, Ir, E_c, F_1}
Where MW_c = Web Crawler
PP_r =Preprocessing
I_r= Itemset Generation and Recursive learning
E_c= Eclat
F_1= Fuzzy Logic
IX. PR = {U_n, MW_c, PP_r, I_r, E_c, F_1, I}

"union of all subset of S Gives the final proposed system".

4. Research Results and Evaluation

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. absolute system performance is calculated by recall.

Consider following equation for computation:
- M = recognized number of relevant Frequent patterns
- N = Not Recognized number of relevant Frequent patterns
- O =Recognized number of irrelevant Frequent patterns .

Then

$$Precision = \left(\frac{M}{M} + o\right) * 100 \quad \ldots\ldots\ldots\ldots\ldots(2)$$

$$Recall = \left(\frac{M}{M+o} * 100 \ldots\ldots\ldots\ldots\ldots\ldots(3)\right.$$

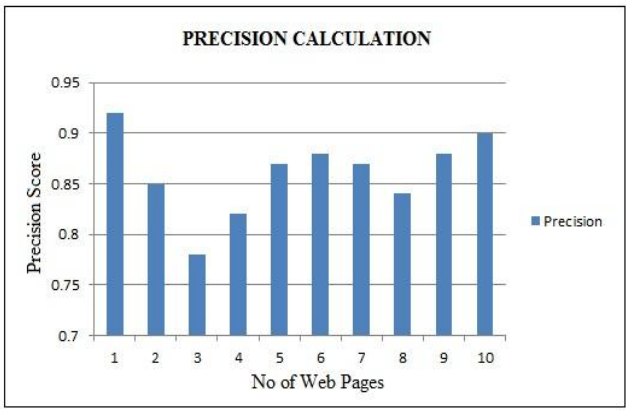
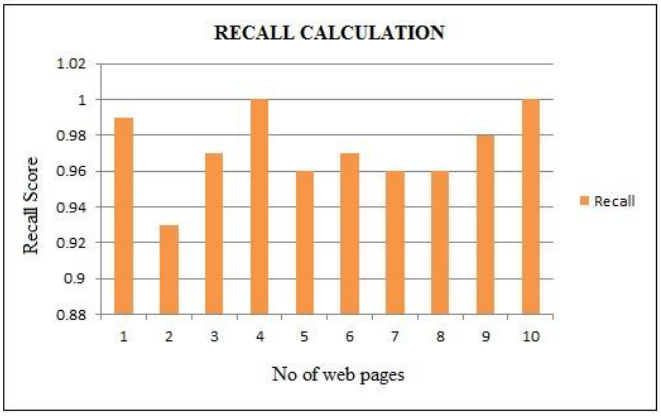


Figure 4: Average precision of the proposed approach

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 3evaluations is 0.85 which evaluates that system recognizes better interesting patterns.



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

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

5. Conclusion and Future Scope

Proposed system sufficiently shows the better precision for the extraction of interesting patterns form the web pages. System efficiently extracts the web pages textual data and parses them too 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 itemsets with their candidates sets and this horizontal data is been converted into vertical data for Éclat mining algorithm. The results of the Éclat is been classified using Efficient rules of Fuzzy logic to identify Interesting patterns of the extracted web data.

Above research can be extended by researchers for distributed information retrieval and finding new better interesting patters.

References

[1] Monica peshave, kamyar dezhgosha "how search engine works and a web crawler application" http://www.micsymposium.org/mics_2005/papers/paper89.pdf [online]
[2] Trupti V.Udapure, RavindraD.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] http://www.worldwidewebsize.com/[online].
[4] http://ausweb.scu.edu.au/aw04/papers/refereed/shokouhi/paper.html[online]
[5] Mitali Srivastava, Rakhi Garg, P. K. Mishra, "Preprocessing Techniques in Web Usage Mining: A Survey" International Journal of Computer Applications (0975-8887) Volume 97–No.1 8July 2014.
[6] Vijayashri Losarwar, Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16,2012 Singapore.
[7] Florez, German, Susan M. Bridges, and Rayford B. Vaughn. "An improved algorithm for fuzzy data mining for intrusion detection." Fuzzy Information Processing Society, 2002. Proceedings. NAFIPS. 2002 Annual Meeting of the North American. IEEE, 2002.
[8] Slimani, Thabet, and Amor Lazzez. "Efficient Analysis of Pattern and Association Rule Mining Approaches." arXiv preprint arXiv:1402.2892 (2014).
[9] Florez, German, Susan M. Bridges, and Rayford B. Vaughn. "An improved algorithm for fuzzy data mining for intrusion detection." Fuzzy Information Processing Society, 2002. Proceedings. NAFIPS. 2002 Annual Meeting of the North American. IEEE, 2002
[10] KumarRana, Ram, and Nidhi Tyagi. "A novel architecture of ontology-based semantic web crawler." International Journal of Computer Applications 44.18 (2012): 31-36.
[11] Khurana, Dhiraj, and Satish Kumar. "Web Crawler: A Review." IJCSMS International Journal of Computer Science & Management Studies 12.01 (2012).
[12] 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.

Author Profile

Sonali Abhane, is currently pursuing M.E (Computer) from Department of Computer Engineering, Jayawantrao Sawant College of Engineering,Pune, India. Savitribai Phule, Pune University, Pune, Maharashtra,India -411007. She received her B.E. (Computer) Degree from MKSSS Cummins College of Engg. For Women, Savitribai Phule Pune University, Pune, Maharashtra, India - 411007. Her area of interest is programming languages & data mining.

Prof. P.D.Lambhate, received her Degree from WIT, solapur, ME(Comp) from BVCOE Pune, Pursing PhD. In computer Engineering. She is currently working as Professor at Department of Computer and IT , Jayawantrao Sawant College of Engineering, Hadapsar, Pune, India 411028, affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India -411007. Her area of interest is Data mining, search engine