ETCOT: A Novel Architecture for Proficient Web Information Retrieval

A.V. Seetha Lakshmi
Assistant Professor of Department of IT
G.T. N Arts College
Dindigul, India
avseethalakshmimphil@gmail.com

Dr. S. P. Victor, MCA, ME, Ph.D
Dean of Science and Associate Professor of Computer Science
St. Xavier's College,
Tirunelveli, India
drspvictor@gmail.com

Abstract- **The web based data retrieval has to interpret the content regarding the Keywords. The day to day effort that leads to huge collection of information through the web. The IR systems help to retrieve necessary information from massive databases over the internet. The key concern of this paper is to reduce the number of assessments and it will reduce the time consuming and provide the optimized search result. The major difficulty would be resolved using a novel architecture known to be Enhanced Theme Condensation and Optimization Technique (ETCOT) for information retrieval. This technique deals with optimizing the content of the web to reduce the assessment analysis and the key objective of this scheme is to obtain the required information from IR System using the categorization technique which summarizes the contents of the web into a normalized form. The assessment analysis is minimized to overhead the content which occurs in various web pages, this kind of dispensation is acknowledged theme condensation. Thus the new technique is implemented in this context which deals with these IR systems in formulated manner and provides an optimized search outcome.**

Keywords- Content Data Mining, Theme summarization, Optimized Theme, Theme condensation, Domain based Association, Optimized Dataset, Clustering, Clustered Data

I. INTRODUCTION

Data mining is the investigation and study of huge data sets; in regulate to discern significant pattern and rules. The intention of data mining is considered for, and effort preeminent with big data sets. [1] It is having assorted statistics of technique which have several possess qualifications, but in this paper, we will concentrate on clustering techniques and its methods. Thus, the users need an optimized and summarized outcome as their searching result. The existing technologies of the search engine or any other searching technique was fetched the conclusions in dissimilar manner similarly adapted web search. In this planned architecture the description technique of summarizing the dataset to obtain the theme condensed data as the user's result. The outcome of this technique is to provide current theme of the webpage to reduce the user's inconvenience of evaluating the theme.

Web Content Mining Techniques

The "Web Content Mining" engages techniques for shortening, organization and clustering of the web contents [2]. It can afford functional and appealing patterns about user desires and involvement activities. It is essentially based on examine in information retrieval and content mining, such as information removal, text categorization and clustering, and information visualization [2]. The web content mining techniques are as follows:

A. *Unstructured Data Mining*-In this technique, the information is explored and repossess includes the text format web pages.
B. *Structured data mining*-This technique is an enhanced scheme for extorting the data from the data retrieval system called Wrapper.

C. *Semi-structured data mining*-Semi-structured data is a tip of union for the Web and database neighborhood: the former deals with documents, the latter with data. HTML is a special case of such "intra document‟ structure.

D. *Multimedia data mining*-The objective for doing Multimedia data mining is to employ the exposed patterns to progress decision making.

Uses of Web Content Mining

The uses of web content mining are as follows:

- To determine the relevance of the content to the search query.
- Improve the navigation of information on the web provides productive promotion.
- Produce a higher quality of information to the user.

The major problem of searching a web for the user is that they have to interpret the entire content of the resultant web pages [2] [3]. The suitable storage space of web documents directs to the complicated in exploration while there are progression of web pages available by various novelist for the particular keyword. Appropriate to the exceptional development of a huge amount of documents in the web it is intricate for the users to understand the complete contents and conclude the accurate theme in the current document. This makes the situation even worse due to time consumption [3]. In this work we identify a assignment which summarizes the interior content in the sequential order and this formulates to readers to recognize the content easily This Context focuses on the subject summarizing task for the user handiness of interpreting regarding the particular topic from the web pages. Thus in this paper a novel technique for the theme modeling and to attain an optimized set of web URL's to make the user significance in penetrating the web.

II. Motivation

In general, data mining reveals motivating patterns and associations concealed in a huge volume of raw data, and the consequences valve out may help formulate important prophecy or prospect annotations in the existent world. Not only the scale of data generated today is unprecedented [3], the created data is frequently always generating in the outline of brooks that necessitate being processed and mined in real time. Delayed detection of still extremely important information quashes the worth of the exposed information.

Information retrieval not only carries new disputes, but also conveys prospects – the unified data with intricate and heterogeneous contents tolerate new resource of information and approaching. In the paper [4], Data would grow to be a useless massive if we don't encompass the accurate utensils to harness its wildness. In progress data mining techniques and algorithms are not prepared to convene the new disputes of data. Mining data demands highly scalable strategies and algorithms, more effective pre processing.

Some of the inadequacies of the data mining regarding the modern technology are as follows:

- *Redundant Data:* Data is exact valuable entity regarding the user's exploration in the information retrieval system. Every user in the web suppose for an exact and accurate.
- *Misuse of information/inaccurate information:* Information is serene during data mining proposed for the decent intentions can be distorted. This information may be broken by disreputable inhabitants or businesses to take remuneration of vulnerable group or differentiate adjacent to a group of people.
- *Security issues: Security is a big issue. There include be a bundle of cases that intruders admittance and wrap large data of regulars from big business with so much personal and financial information available.*

Data mining brings a lot of benefits to businesses, society. Although confidentiality, security and abuse of information are the big harms if they are not deal with and determined appropriately.

III. Problem Statement

In existing techniques **[**5**]**, the user have to explore for the require information in retrieval system is performed under *Keyword-based Association Scheme for Theme Condensation (KASTC)*. In KASTC, the search word specified by the user is evaluated in the form of keyword extracted from the user requirement and provides the keyword based outcomes. The outcomes complicate the users to carry out their constraint from the huge amount of resultant web documents. In additional to this **[**6**]**, KASTC would retrieve the information regarding the keyword and the also retrieve the related web documents based on keyword similarity in the web documents.

From the massive quantity of web documents the user could carry out their requirement, this would construct a major complexity and takes lot of time even for single information.

The key limitations for the existing technique KASTC,

- Retrieves the information based on the specified keyword only.
- Carries out more distinct outcomes for user's particular search
- Predicts the web pages related to the keywords based on document similarity.
- Keywords based document similarity leads with a great amount of redundant information or web pages.
- This makes more complication in web search and receives much time consumption.

Thus in this paper, we propose a novel technique to make ease for the user to obtain the core content as their exact and precise outcome. The novel technique proposed in this study known to be ETCOT (Enhanced Theme Condensation and Optimization Technique). This technique retrieves the content based on the Domain. The Domain-based Association Scheme (DAS) is included in this context to formulate the user's requirement effectively and efficiently.

IV. Proposed Architecture

Established technique of penetrating the web was using contents. Web Content mining is the comprehensive works achieve by Information retrieval. Web Content mining refers to the finding of valuable information from network satisfied [6]. The information retrieval also provides bulk of search outcomes that is unfeasible for abuser to go through each.

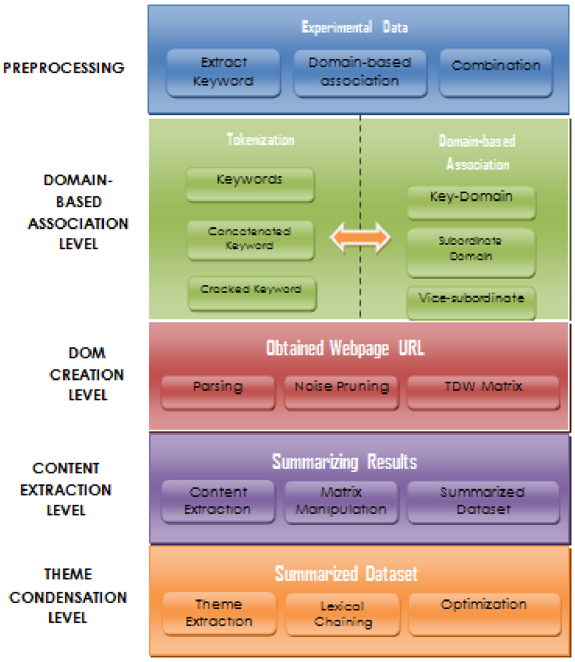


Figure 1. Enhanced Theme Condensation and Optimization Technique. *(ETCOT)*

Thus in this context, the content mining is handled with the specified manner that would summarizes the similar content mined from the multiple web-pages and the theme of that particular results would be summarized in a format for the user's expediency.

This architecture classified into five main phases:

a) Pre-processing
b) Domain-based Association Scheme (DAS)
c) DOM Creation
d) Content Extraction
e) *Theme Modeling*

With the increasing digital data repositories and the demand of data centric research in data mining society, finding suitable dataset for a examine trouble has happen to an essential step in scientific research. But specified the extensive assortment of data convention in scientific research it is very tricky to figure out which datasets are most practical for a fastidious investigate theme [7] [8]. To improve this trouble, a mechanized dataset search engine is an influential implement.

However, receiving information about dataset handling involves a keyword search or physically going through the details of the works that have used the datasets [8]. The ordinary data preprocessing such as stop words removal, tokenization and HTML tag exclusion are pertained to the keyword k and tokens {x1, x2, x3…..xn}. Finally, the weighting scheme W is included regarding the term x for the input keyword.

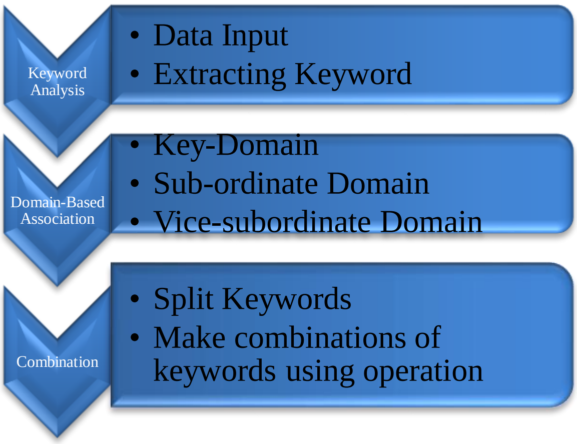


Figure 2. Preprocessing

A. Preprocessing

In the first phase, user defined is keyword is extracted for the **preprocessing**. The Fig.2 depicts the process flow of preprocessing. The extracted keywords are searched in the form of novel technique known to be ***Stern-Probing***. The stern-probing technique in this context is involved in penetrating the extracted keywords in the type of Domain-based Association Scheme (DAS). This DAS brings out the perfect and the associated outcomes for the user's keyword. The stern-probing, approaches the user with a superfluous mining regarding their searched keyword, it provides an additional deposit as result for the further processing and it would reduce the complexity of searching from the huge dataset of web.

B. Domain-based Association Scheme(DAS)

The second phase of this architecture is to make domain-based association with the extracted keywords. The extracted keywords are shambled and concatenated to make the domain-based association factor. Then the combinations for the keywords would be produced for the domain-based factor. The combinations of keywords are conceded for the progression of *Tokenization*. The process of tokenization is to progress combined keywords with the concatenated operators to the process of tokenizing every word for the extraction of URL.

C. DOM Creation

In the third phase, the efficiency of feature extraction and finally classification accuracy are certainly degraded due to the occurrence of such piercing information [14]. Thus clear out the web pages prior to mining develops into serious for civilizing the mining consequences [9]. In this effort, spotlights on recognizing and eliminating restricted noises in web pages to develop the recital of mining. This paper proposes a novel and simple idea for the discovery and exclusion of confined resonances using a new hierarchy construction called DOM Tree. Noise removal can be executed as a pre-processing pace for content mining.

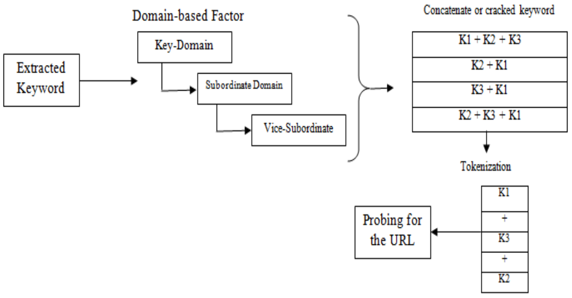


Figure 3. Domain-based Association Scheme *(DAS)*

D. Content Extraction

It is projected [14] by merging a dissimilar term weighting approach for finest characteristic subset selection, DOM tree modeling of the complete page presenting the description.

The Fig.3 shown above describes the *stern-probing* approach to penetrate the extracted keyword to make search in the form of domain-based association. The stern-probing technique operates as a *back-linking technique*, i.e. the keyword is associated with domain based dataset in the server. This domain-based association makes the probing progress efficiently and precise outcomes for searched keywords. Then the keywords are cracked and concatenated based on the domain in the probing technique. Finally, the split or concatenated keywords are approved to the progression of Tokenization for the probing the URL. [10] If a user request to observation a fastidious page along with server log entrance graphics and scripts are download in accumulation to the HTML file. To summarize the content extraction complexity the occurrence of the keywords in the content words are evaluated to represent a matrix [11]. The matrix representation is evaluated to minimize the complication involved in the theme extraction for the user expediency.

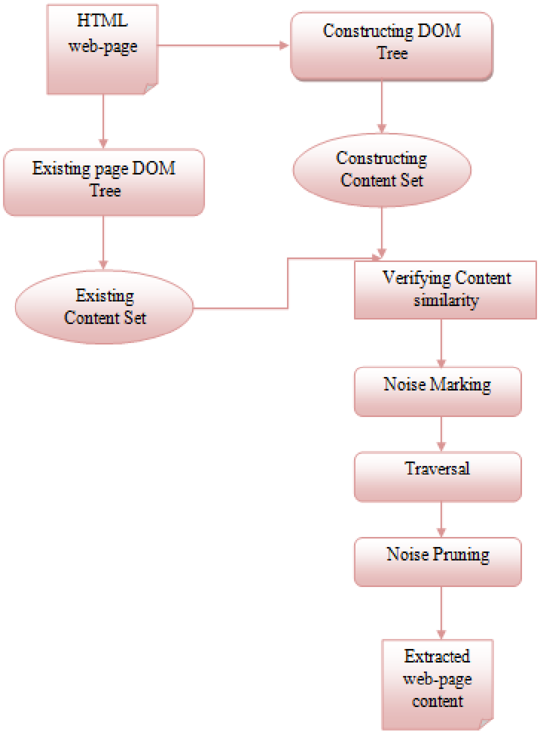


Figure 4. DOM Creation

In this context, the TDW (Term Document Weight) Matrix is used characterize the related keywords from the content of web-pages [12]. The huge amount of web documents has resulted in tribulations for information retrieval important to the reality that the search consequences are of less significance to the user. In this paper, we propose a novel idea for finding near duplicates of an input web-page, from a enormous warehouse.

TABLE I. TERM-WEIGHTING

Term	Weighting
Head	1
Body	3
URL	1
Paragraph	2
Heading	1
Anchor tag	1
Keywords	3
Description	3

This advance investigates the semantic structure, content and context, of a web page quite than the content only approach. The weighting scheme recommended in is measured for generating a TDW which plays [15] a significant responsibility in the proposed algorithm. We present a three-stage process which obtains an input record and a threshold value and returns an optimal set of data.

TABLE II. SAMPLE TDW

Terms\Records	r1	r2	r3
t1	0	$\frac{Wt1r2}{\sum \mathrm{Wtir2}}$	$\frac{Wt1r2}{\sum \mathrm{Wtir2}}$
t2	$\frac{Wt2r1}{\sum \mathrm{Wtir1}}$	$\frac{Wt2r2}{\sum \mathrm{Wtir2}}$	$\frac{Wt2r3}{\sum \mathrm{Wtir3}}$
t3	0	$\frac{Wt3r2}{\sum \mathrm{Wtir2}}$	0
t4	$\frac{Wt4r1}{\sum \mathrm{Wtir1}}$	0	$\frac{Wt4r3}{\sum \mathrm{Wtir3}}$

Here this paper propose an innovative idea for finding nearest web pages of an input record r. Similarity confirmation is completed on a huge record set having n number of records {r1 ,r2 ,….,rn} and an optimal set of similar records are arrival. Our purpose is to discover how to decide on related records from the complete record set with a condensed number of evaluations. Each page thus obtained is pre-processed, attributed and weighted according to the weighting system and accurately indexed to create a TDW matrix.

TABLE III. RESULTANT MATRIX

Records	Obtained URL				
r1	url3	url2	url3	url5	-
r2	url5	url1	url3	url5	url2
r3	url3	url1	url4	url2	url1
r4	url2	url1	url4	url3	url5
r5	url1	url3	url2	url1	url3

If a term x is present in a record r, its total weight is represented as Wxr. The document occurrence of a token is the number of records that enclose the token. Let r1, r2, r3 be three records. r1={x2, x1, x3} r2={x4, x1, x3} r3={x2, x4, x1, x3} The TDW matrix is given in figure 12, where 1 ≤ i ≤ 4. Prefix length of r is calculated as:

$$\text{Prefix length} = |r| - \lceil t.|r| \rceil + 1 \qquad (1)$$

Assume that the similarity threshold value is ***t***, and then each term of the record in prefix set ***r*** is related with all records prefix set $\mathbf{P_s}$ in the dataset [15]. If there is any record **ri** is distributing a term with **r**, then it is added to pre-final set $\mathbf{P_{fs}}$. The record can be avoided from the other processing, if there is no terms are familiar in the prefix set. In order to shorten the unrelated records from the final prefix-set the prefix filtering and positional filtering should be pertained.

E. Theme Modeling

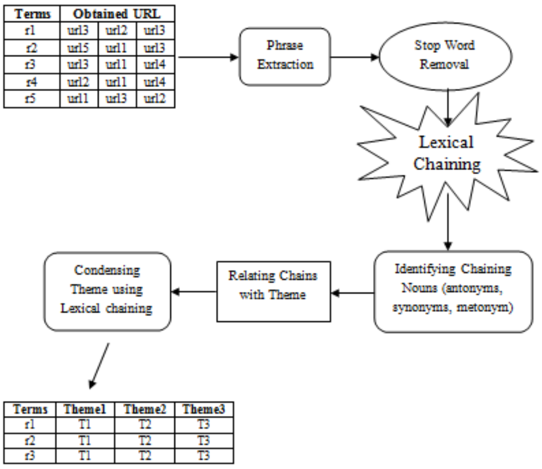
In the opinion mining progression, topics are noun phrases with related significance attains. Themes extract accurately as illustrated in noun phrase extraction [16] [17]. Once extorted, themes are then achieved for appropriate relevance using **lexical chaining**. Initially, probable themes are extracted support on the part-of-speech patterns. Then the chains attained and the themes with highest-scoring chain together. If there are less than four chains, the algorithm elegantly corrupts to getting merely by count. [18] With the theme extraction, their scores would be diverse depending on where they raised in the text.

Algorithm:

- K-Input Keyword
- D-Keyword Dataset
- C- Keyword Combination
- W-Extracted Web Pages
- TDW-Term-Document Weight Matrix
- T-Similar Terms in document
- S-Removal of Stop words in Theme
- N-Extracted Themes

```
Algorithm ETCOT (Keyword K)
    Extract $K'$ for each Keyword K
    Check (Domain ($K'$) from D Dataset)
    If (Domain ($K'$) Exists) then
    Tokenize ($K'$)
    C = Combine ($K'$) for Probing
    W = Search (C)
    DOM CREATION ($W'$) for each Web pages W
    STRENPROBING ($W'$) from Back-Linking
    Extract Content from $W'$
    Represent TDW for Term-Weighting
    TDW = Split ($T'$) of $W'$
    Theme Extraction ($W'$) for each W
    If (N of $W'$) then
    Summarize the Theme S by Stop words removal
    Optimize (N) // the Result by Clustering
End Algorithm
```

Lexical Chaining is an imperative procedure in expected idiom dispensation; several of Lexalytics algorithms rely greatly on it. Lexical Chaining narrates stretches using thesaurus-oriented nouns. Even if those sentences are not contiguous to each one in the manuscript, [19] they are lexically associated to each other and can thus be linked with each other. This is a actually essential notion - if the nouns are correlated to each other, we can locate that lexical chain in the substance, even when those sentences are divided by many other isolated sentences. The gain of a lexical chain is straightforwardly interrelated to the span of the chain and the relationships among the chaining nouns such as identical word, antonym, synonym, meronym, hyper/homonyms, etc.



Terms	Obtained URL		
t1	url3	url2	url3
t2	url5	url1	url3
t3	url3	url1	url4
t4	url2	url1	url4
t5	url1	url3	url2

Phrase Extraction
Stop Word Removal
Lexical Chaining
Identifying Chaining Nouns (antonyms, synonyms, metonym)
Relating Chains with Theme
Condensing Theme using Lexical chaining

Terms	Theme1	Theme2	Theme3
t1	T1	T2	T3
t2	T1	T2	T3
t3	T1	T2	T3

Figure 5. Theme Modeling

Optimization using Clustering Technique

Clustering is an essential chore of examining data analysis and data mining relevancies. Clustering is the charge of combination a set of objects in such a way that objects in the similar collection called a cluster are further related to each other than to those in other clusters [20]. The generally used approaches are in Clustering are Hierarchical, Partitioning based algorithms. In this system we crack the information into cluster and these collections are known as clusters. A data desires the cluster regarding to quality standards recitation by objects [21]. Cluster Central will symbolize with input vector can tell which cluster this vector fit in to by determining a resemblance metric involving input vector and all cluster centers and decisive which cluster is adjacent or most related one.

A. HandyRanking Clustering Algorithm

HandyRanking clustering is a centric supported technique. This technique affords the tree association among clusters. In this process we employ similar no. cluster and information, resources if we have‘d’ no. of data then we use k no of clusters. HandyRanking clustering constructs a hierarchical disintegration of the set of data via various conditions. It can be envisioned as a fact that is a tree similar to illustration that records the progressions of combines or splits. Any preferred number of cluster can be attained by removing the data at the accurate stage. [22] Each cluster node includes adolescent clusters; sibling clusters separation the positions enclosed by their familiar parent. This technique receives the input bounds k and separation a set of n objects into k clusters that the resultant intra-cluster relationship is elevated but the inter-cluster relationship is low down [23]. The technique can be used by cluster to allocate grade ideals to the cluster unqualified data is arithmetical technique. Nearest is essentially based on the expanse among the object and the cluster suggest. Then it calculates the original mean for every cluster. Here unconditional information has been transformed into numeric by transmission grade value.

Algorithm:

- K- The number of cluster
- D- The data set containing an object.
- T-Theme Set
- C-Similar themes related to theme
- N-Optimized Set of Theme

```
Algorithm HANDYRANKING (T as Tree, K)
    Parse ($T'$) from T as Tree
    Relation(C) from Tree $T'$
    Analyze ($C'$) of C where Children or Siblings
    If ($C'$ Exists) then
    Split or Combine ($C'$) to $T'$ to Construct Cluster
    Construct (N) = Combine ($C' + T'$)
    Random (N) from D as initial Cluster
    Calculate (D) = Distance ($N'$) from Data Point to each Cluster
    If ($\sum D \neq \sum N'$) then
    Move to the closest cluster $N'$ for Theme
    Repeat for each data $N'$
    Mean = ($\sum D . \sum N'$) / N
    Locate ($N'$) to the position for accurate theme
    Construct Theme N
    Optimize Set N
End Algorithm
```

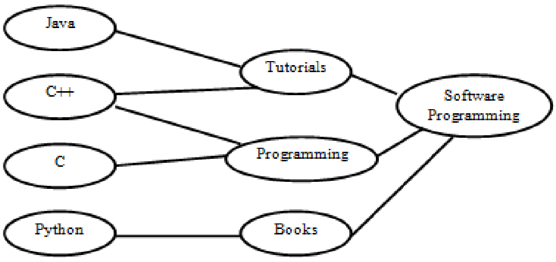


Figure 6. HandyRanking Clustering

This approach permits discovering data on dissimilar stages of granularity. Hierarchical clustering are classifying into bottom-up and top-down. A bottom-up clustering creates with singleton clusters and recursively combines two or more of the most related clusters. Top-down clustering creates with a sole cluster that holds all data points and recursively cracks the most suitable cluster. The procedure repeats until an ending condition frequently, the requested number k of clusters is attained.

V. Performance Evaluation

The projected design of ETCOT is employed using Asp with C#.Net as web-based system in visual studio 2012, NET framework 4.5 and SQL Server 2012 and implemented on a Windows 8.1, 64 bit system setting with an Intel Core i5-2410M CPU@2.30 GHz with 4 GB RAM and 500 GB hard disk. The assessment information is the web-pages composed together from various domains such as Education, Software's and Bio-information Data. This framework has created an enormous depository of web page records from information retrieval system. It was completed for some detailed inquiry words given to IR and composed all comparable pages with deference to the first graded result of IR. Some of these records were removed due to the lack of essential contents retrieved and non-suitable file formats like PDF. Each page thus gained is pre-processed, characteristic and weighted regarding to the weighting scheme and appropriately indexed to generate a TDW matrix.

TABLE IV. RESULTS OBTAINED URL

Records	Obtained URL			
Java	www.tutorialspoint.com	www.tutorialspoint.com	www.java-examples.com	-
Java program	www.tutorialspoint.com	www3.ntu.edu.sg	www.cs.usfca.edu	-
Java programming	www.programmingsmiplified.com	-	-	-
Programming	www.tutorialspoint.com	www.programmingsmiplified.com	www.oracle.com	-
Basics	www3.ntu.edu.sg	www.javabeginnerstutorial.com	www.java-examples.com	-

The efficiency of the data retrieval system is frequently considered by the proportion Precision, Recall and Average Precision. The highly similar web pages are retrieved using Precision value P. The capability of retrieving all the related web documents from the massive quantity of data is Recall R.

P=RD/N
RD-No. Of similar documents retrieved
N-Total no of documents
R=RD/T
RD-No. Of similar documents retrieved
T- Total no of similar documents

TABLE V. PRECISION AND RECALL FOR KEYWORD COMBINATIONS

Terms	ETCOT		KASTC	
	Precision	Recall	Precision	Recall
K1	0.86	0.74	0.32	0.23
K2	0.75	0.56	0.23	0.5
K3	1	1	0.41	0.12
K4	0.82	1	0.49	0.25
K5	1	0.89	0.21	0.32
K6	0.74	0.65	0.47	0.28
K7	0.68	0.89	0.51	0.36

TABLE VI. EVALUATION TABLE

Evaluation Metrics	Proposed Method (ETCOT)	KASTC
Keyword Analysis	Based on the relation between the keywords using DAS	Based on the meaning of the keywords
Probing	Domain based association Search	Keyword association Search
Content Extraction	DOM based Extraction	Block-wise Extraction
Theme Extraction	Summarizing theme based on Lexical Chaining & DAS	Keyword based theme extraction
Accuracy	Provides Highly accurate information for keyword	Lack of accuracy because only based on keyword

This process was frequent for various dissimilar queries and trials were performed on some unusual repositories thus produced. Each trial is resulted an optimal set of web pages with esteem to the high-graded web page in the query result.

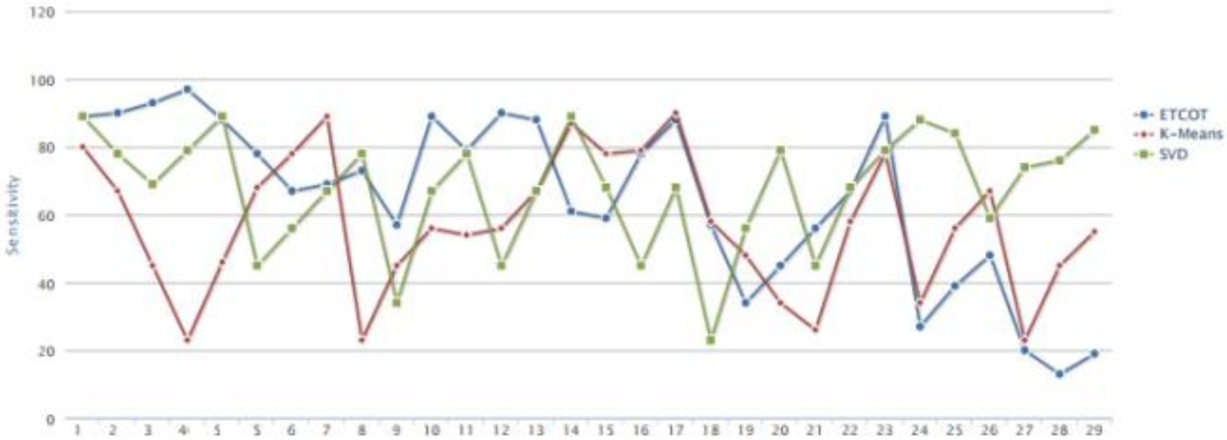


Figure 7. Sensitivity Analysis

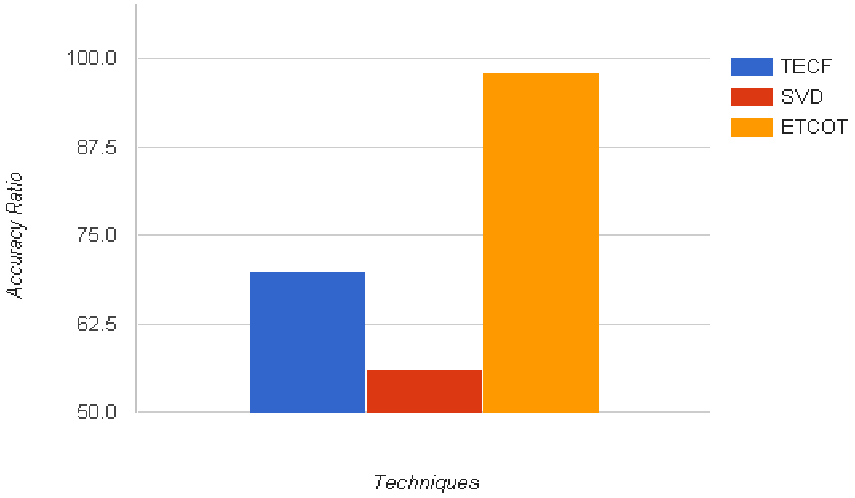


Figure 8. Theme Extraction

To accomplish essential trials, created framework which retrieves the contents of an input web page; achieve all pre-processing paces and extract weighted feature set.

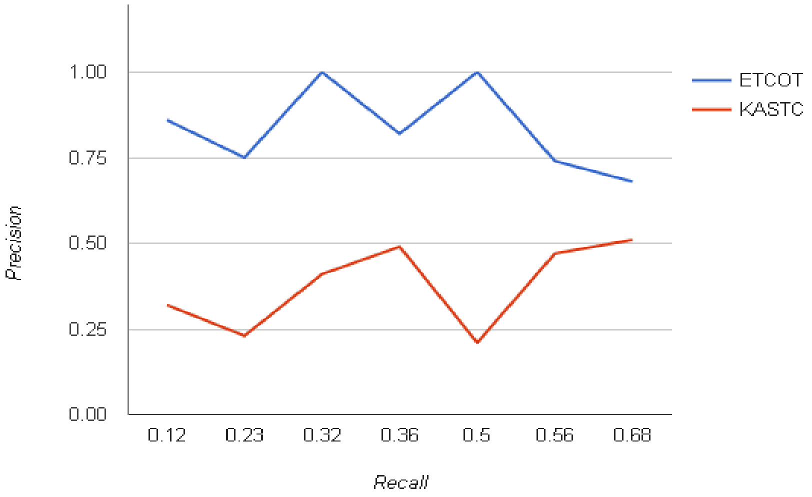


Figure 9. Average Precision

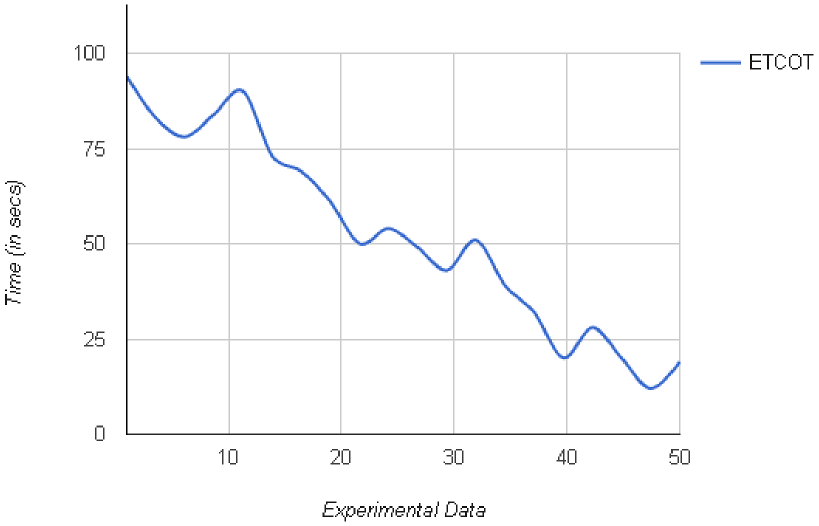


Figure 10. Performance Chart

With respect to the large warehouse generated previous, a TDW matrix is formed in a universal arranging and straining ethics are applied.

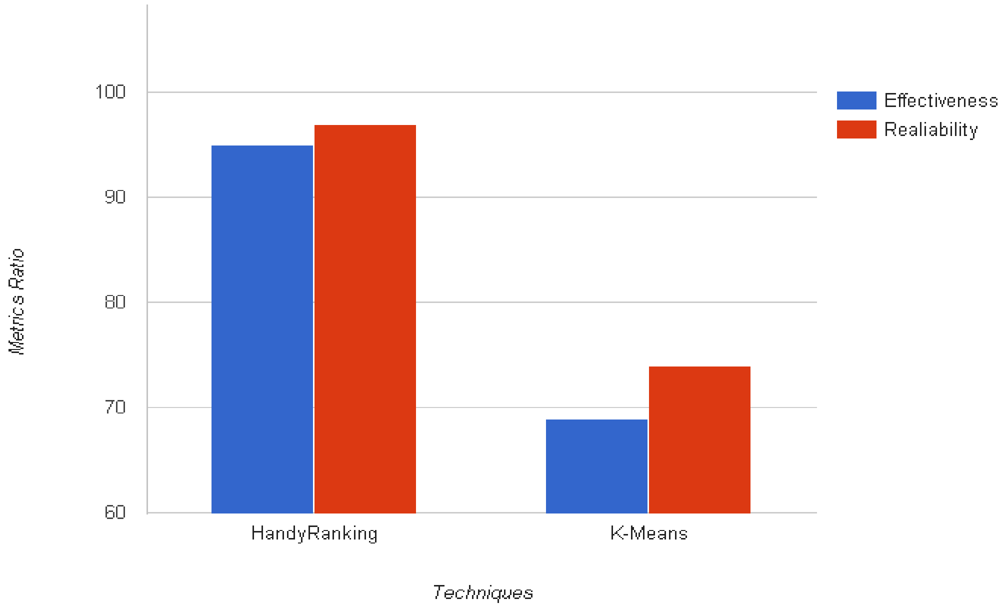


Figure 11. Effectiveness and Realability

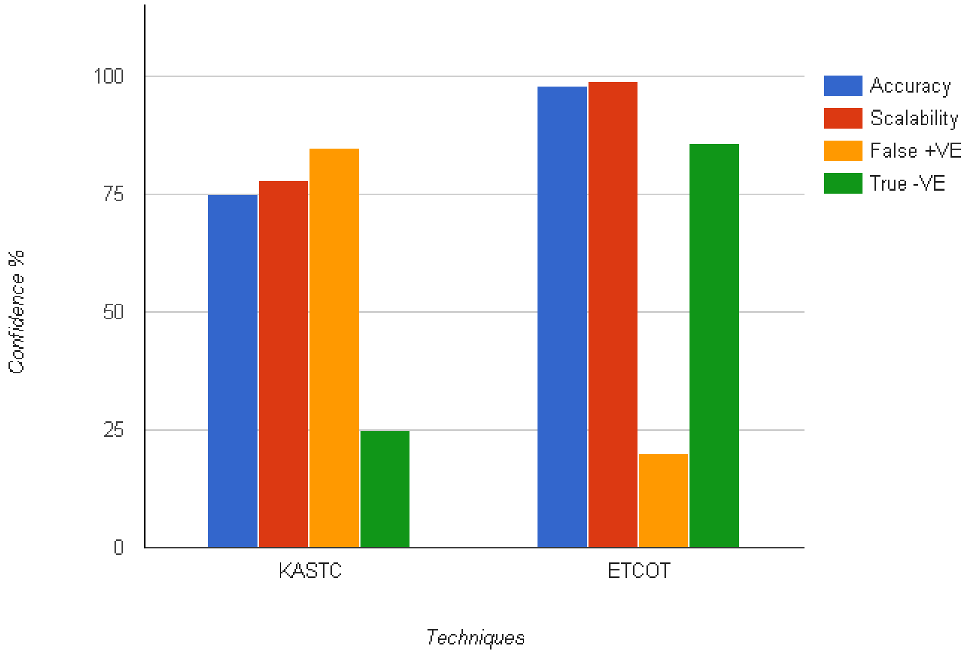


Figure 12. Confidence Level

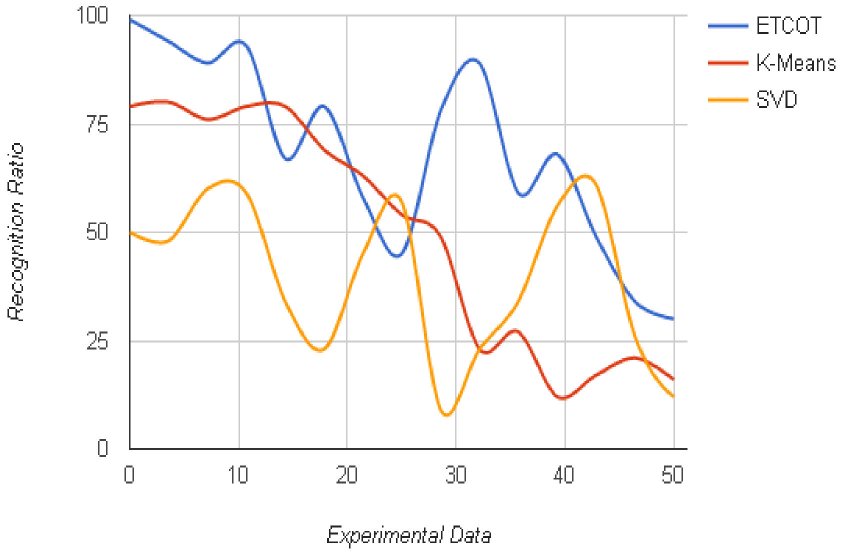


Figure 13. Pattern Recognition

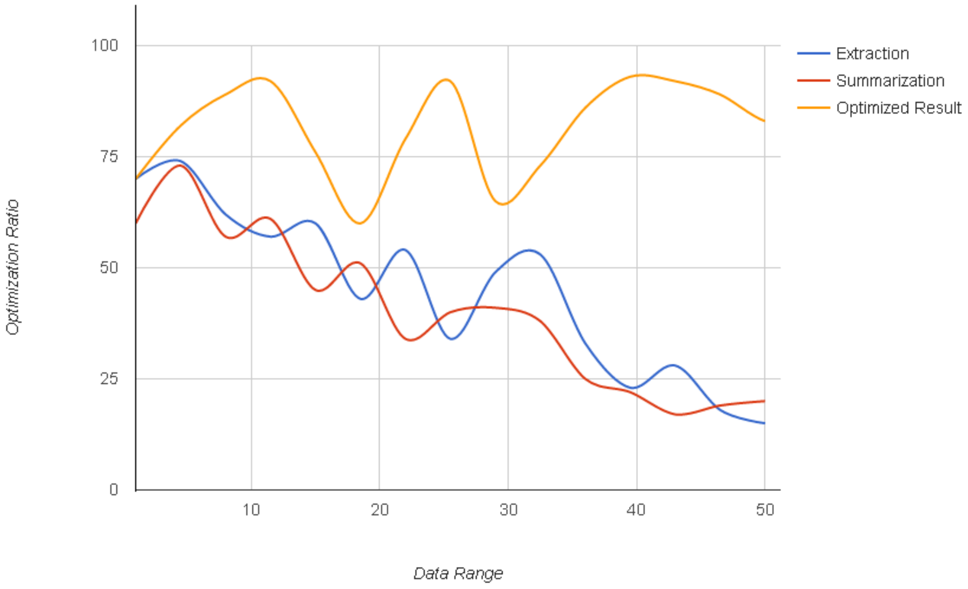


Figure 14. Optimization Chart

TABLE VII. EVALUATION TABLE

Keywords	Theme1	Theme2	Theme3
Java Program	High-level programming	Sun Microsystems	Object oriented
	OOPS Concepts	Collection of objects	Robust
	Objects and Classes	Basic data types	Multithreading
	Java basic programs	Environment setup	Dynamic
Java Programming	Elements of Programming	Example programming	String concepts
	Built-in data types	Compiling and executing	Basic concepts
	Conditions and looping	Development IDE	Arrays
	Input and output statements	Environments	Java Kit
Java	Java software	Programming Language	Versions of java
	Focus on games	High-level programming	Java applets
	Java Plug-in	Improves application service	JVM
	Runtime environment	System requirement for java	Java runtime

The resultant matrix thus gained is completed for correspondence confirmation by examining of each record with the input record. If it assures the threshold value t, that record is noticeable as an optimized one.

VI. Conclusion and Future Enhancement

This paper concluded that the problem of finding the exact contents of the resultant web pages related to their survey the user would obtain their demand contented outcome as their requirement. This anticipated technique ETCOT, proposes to assisting classification of content mining based on the Theme condensation and optimization schemes. As an alternative of using customary similarity and optimization this paper focused to reduce the volume and enhances the relevancy for obtaining the condensed theme as the user need. This technique provides the correctness and scalability in name of precision and recall. The performance results would provide a better outcome when compared.

In the future work, the summarization of the content may hold in very undersized form of outcome for the user access requirement. It will be more efficient and consumes a lesser amount of time to observe the content for the particular concept. This system may increase the work for well-organized approach for better finding of results.

References

[1] S.Nagaparameshwara Chary, “A Study on Fundamental Concepts of Data Mining”, International Journal of Computer Science And Technology, Vol. 2, Issue 4, Oct. - Dec. 2011

[2] Faustina Johnson Santosh Kumar Gupta, “Web Content Mining Techniques: A Survey”, International Journal of Computer Applications, Vol 47 No 11, 2012.

[3] Dr.S. Vijiyarani, Ms. E. Suganya, “Research Issues in Web Mining”, International Journal of Computer-Aided Technologies, Vol.2, No.3, July 2015

[4] Md. Zahid Hasan, Khawja Jakaria Ahmad Chisty and Nur-E-Zaman Ayshik, “Research Challenges in Web Data Mining”, International Journal of Computer Science and Telecommunications,Vol 3, Issue 7, July 2012

[5] Ananthi.J, “A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites”, International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014

[6] Tae-Hoon Lee, Jung-Hyun Kim, Hyeong-Joon Kwon and Kwang-Seok Hong , “Keyword-based Semantic Retrieval System using Location Information in a Mobile Environment”, Proceedings of the 2009 International Symposium on Web Information Systems and Applications, , May 22-24, 2009.

[7] Ambesh Negi, Mayur Bhirud, Dr. Suresh Jain, Mr.Amit Mittal, “Index based Information Retrieval System”, International Journal of Modern Engineering Research, Vol.2, Issue.3, May-June 2012

[8] Broder, A., Glassman, S., Manasse, M., And Zweig G. “Syntactic Clustering Of The Web”, In 6th International World Wide Web Conference, Pp: 393-404, 1997.

[9] Fetterly, D., Manasse, M. And Najork, M. “On The Evolution Of Clusters Of Near Duplicate Web Pages”, In Proceedings Of The First Latin American Web Congress (Laweb), 37–45, 2003.

[10] Yun Ling, Xiaobo Tao Hexin Lv, “A Priority-Based Method Of Near Duplicated Text Information Of Web Pages Deletion”, IEEE International Conference On Software Engineering And Service Sciences (ICSESS), August 2010.

[11] Fatiha Boubekeur and Wassila Azzoug, “Concept-Based Indexing In Text Information Retrieval”, International Journal of Computer Science & Information Technology, Vol 5, No 1, February 2013

[12] Arun P. R, Sumesh M. S, “An Extended TDW Scheme by Word Mapping Technique”, International Journal of Engineering Research & Technology, Volume. 4 - Issue. 07 , July - 2015

[13] Faustina Johnson And Santosh Kumar Gupta, “Web Content Mining Techniques: A Survey”. International Journal Of Computer Applications (0975 – 888) Volume 47– No.11, June 2012

[14] S. Mythili, T. Vetriselvi, “Analytics of Noisy Data in Web Documents Using a Dom Tree”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, April 2015

[15] Arun P R, Sumesh M S, Eldhose P Sim, “Web Page Categorization with Extended TDW Scheme”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Special Issue 1, June 2015

[16] V. Bharanipriya & V. Kamakshi Prasad, “Web Content Mining Tools: A Comparative Study”.

[17] Chidansh Amitkumar Bhatt, Mohan S. Kankanhalli, “Multimedia Data Mining: State Of The Art And Challenge”, Journal Multimedia Tools And Applications Archive Volume 51 Issue 1, January 2011.

[18] Syed Salman Ahmed, Zahid Halim, Rauf Baig, And Shariq Bashir, “Web Content Mining: A Solution To Consumer‟S Product Hunt”, International Journal Of Social And Human Sciences 2 2008.

[19] Faustina Johnson And Santosh Kumar Gupta, “Web Content Mining Using Genetic Algorithm”, Advances In Computing, Communication, And Control Communications In Computer And Information Science Volume 361, 2013, Pp 82-93

[20] Aastha Joshi, Rajneet Kaur, “A Review: Comparative Study of Various Clustering Techniques in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013

[21] Ilyinsky, S., Kuzmin, M., Melkov, A., Segalovich, I., “An Efficient Method To Detect Duplicates Of Web Documents With The Use Of Inverted Index”, Proceedings Of The Eleventh International World Wide Web Conference, 2002.

[22] Soto Montalvo Victor Fresno, Raquel Martinez, “Improving Web Page Clustering Through Selecting Appropriate Term Weighting Functions”, 1st International Conference On Digital Information Management ,2006

[23] Gurmeet Singh Manku, Arvind Jain And Anish Das Sarma, “Detecting Near Duplicates For Web Crawling”, Proceedings Of The 16th International Conference On World Wide Web, Pp. 14-150, Ban ,Alberta, Canada, ,2007.

[24] Raquel Martinez. Alberto P.Garcia-Plaza,Victor Fresno, “Web Page Clustering Using Fuzzy Logic Based Representation And Self-Organizing Maps”, IEEE/WIC/ACM International Conference On Web Intelligence And Intelligent Agent Technology ,2008

[25] Xuemin Lin Chuan Xiao, Wei Wang, “Efficient Similarity Joins For Near Duplicate Detection”, 17th International Conference On World Wide Web ,2008

[26] A.V.Seetha lakshmi, Dr.S.P. Victor, “TECF: Accomplishment of Content Framework Tactic to Temporal Theme Condensation”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 11, Nov 2013

A.V. Seetha Lakshmi, Assistant Professor in Department of Information Technology, G.T.N college (Autonomous), Dindugal, received her M.C.A. degree from Madurai Kamaraj University, Madurai, and M.Phil degree from Madurai Kamaraj University, Madurai. She has undertaking her Ph.D. degree in Computer Applications for her research in Data Mining. She is teaching Computer Science since 2005. She has been the Head of the Department of Information Technology for Eight years. Shei has recognized her as a research guide for M.Phil Candidates. So far 3 candidates have completed their M.Phil degree under her guidance. She has published 5 research papers in international journals.

Dr. S. P. Victor, Associate Professor in Computer Science, St. Xavier’s college (Autonomous), Palayamkottai, Tirunelveli, received his M.C.A. degree from Bharathidasan University, Tiruchirappalli, and M.E. (CSE) degree from Anna University, Chennai. The M. S. University, Tirunelveli, has awarded him Ph.D. degree in Computer Applications in the year 2005 for his research in Parallel Algorithms. He is teaching Computer Science since 1988. He has been the Head of the Department of Computer Science and the Director of the Computer Science Research Centre for six years. The M.S. University, Tirunelveli has recognized him as a research guide. So far 12 candidates have completed their Ph.D. degree under his guidance, and 8 candidates are pursuing research. He has published more than 65 research papers in international journals. He has organized Conferences and Seminars at the state and national levels.