

Major Project-II Report
On
**CONTENT BASED RE-RANKING OF WEB
DOCUMENTS USING THE CLUSTERING AND
CLASSIFICATION TECHNIQUES**

Submitted in Partial fulfilment of the Requirement for the Degree of
Master of Technology
in
Computer Science and Engineering

Submitted By
Dhermendra Kumar
2K16/CSE/05

Under the Guidance of
Mr. Sanjay Kumar
(Assistant Professor)



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahabad Daulatpur, Main Bawana Road, Delhi-110042

June 2018

CERTIFICATE

This is to certify that Project Report entitled “**content based re-ranking of web documents using the clustering and classification techniques**” submitted by **Dhermendra Kumar** (2K16/CSE/05) in partial fulfilment of the requirement for the award of degree Master of Technology (Computer Science and Engineering) is a record of the original work carried out by him under my supervision.

Project Guide

Mr. Sanjay Kumar

Assistant Professor

Department of Computer Science & Engineering

Delhi Technological University

DECLARATION

I hereby declare that the Major Project-II work entitled “**content based re-ranking of web documents using the clustering and classification techniques**” which is being submitted to Delhi Technological University, in partial fulfilment of requirements for the award of the degree of Master of Technology (Computer Science and Engineering) is a bona fide report of Major Project-II carried out by me. I have not submitted the matter embodied in this dissertation for the award of any other degree or diploma.

Dhermendra Kumar

2K16/CSE/05

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor Mr. Sanjay Kumar for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided without which the project could not have been a success.

Secondly, I am grateful to Dr. Rajni Jindal, HOD, Computer Science & Engineering Department, DTU for her immense support. I would also like to acknowledge Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not the least I would like to express sincere gratitude to my parents and friends for constantly encouraging me during the completion of work.

Dhermendra Kumar

Roll No – 216/CSE/05

M. Tech(Computer Science & Engineering)

Delhi Technological University

ABSTRACT

We are living in modern age, in today's world we have vast amount of data. In order retrieve the meaningful data (information) and to knowledge discovery we perform various operations. In this internet era search engines plays a key role in information retrieval and organize the relevant data for various purposes. But the data return by a search engines is still debatable because search engines gives us our required data for a user query but also return irrelevant and redundant results. Web content mining and information retrieval is an ample and powerful research area in which retrieval of relevant information from the web resources in a faster and better manner. Web content mining improves the searching process and provides relevant information by eliminating the redundant and irrelevant contents. here unlike the page ranking algorithms we are using weighted (content) based ranking algorithm and comparing the results . further we are also using clustering technique and classification for better results . results shows this way the ranking of web documents or pages gives better performance in terms of precision recall and f-measure.

TABLE OF CONTENTS

CERTIFICATE	i
DECLARATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
List of Figures	vi
List of Abbreviations	vii
CHAPTER 1: INTRODUCTION	8
1.1 Web Crawling Process	8
1.2 Search Engines	9
1.3 Page Rank	11
1.4 Web Mining and It's Architecture	13
1.5 Motivation	15
CHAPTER 2: LITERAURE REVIEW	16
CHAPTER 3: PROPOSED WORK	19
3.1 Problem Statement	19
3.2 Related Work	19
3.3 Used Techniues	20
3.4 Proposed Solution Architecture	23
3.5 Working	24
CHAPTER 4: SIMULATION AND RESULT	
Error! Bookmark not defined.	
CHAPTER 5: CONCLUSION AND FUTURE DIRECTIONS	40
CHAPTER6: REFERENCES	41

List of Figures

Fig. 1. Web Crawling Process	9
Fig. 2. Search Engine Process	10
Fig. 3. Web Mining Architecture	13
Fig. 4. Unordered Data VS Clustered Data Comparision	21
Fig. 5. Distance Functions	24
Fig. 6. Random Sensor network	25
Fig. 7. Architecture of Proposed Solution	27
Fig. 8. USER QUERY RESULTS	33
Fig. 9. EXTRACTED DATA	34
Fig. 10. PRE-PROCESSING RESULTS	35
Fig. 11. FREQUENCY COUNT	36
Fig. 12. CLUSTERING RESULTS	37
Fig. 13. CLASSIFICATION RESULTS	38
Fig. 14. RANKING RESULTS	38
Fig. 15. DATASET SIZE VS TIME(SEC)	39
Fig. 16. PRECISIONCOMPARISIONOFEXISTINGVSPROPOSED	40

List of Abbreviations

1. PRA: Page Ranking Algorithm
2. SE: Search Engine
3. WCP: Web Crawling Process
4. K-NN: K- Nearest Neighbour
5. SK: Strength of Keyword
6. F-Measure: Frequency Measure
7. CBR: Content Based Ranking
8. UBR: Usage Based Ranking
9. LBR: Link Based Ranking
10. SC: Strength of Keyword
11. FP: False Positive
12. TP: True Positive
13. FN: False Negative
14. TN: True Negative

CHAPTER 1: INTRODUCTION

1.1 Web Crawling Process (WCP)

World Wide Web plays a starring role for retrieving user requested information from the web resources. Search engines plays a major role for crawling web content and organize them into result pages so that the user can easily select the requested information by navigating through the pages link. Earlier these information resources were limited and it was also feasible to identify the relevant information directly by the user form search engine results. But when internet era came-

1. Sharing of resources increased and hence we need to develop a technique to rank the web content resources.
2. Different search engines uses different searching techniques and also the different ranking algorithms. Now a days the web content and the resources become dynamic with respect to the user query.

Web crawling process simply means getting the required web resoures based on some user query.

Web crawling can be divided into sub parts-

1. Resource finding- finding the relevant web documents.
2. Information selection and pre- processing- the selection of information from retrieved web documents. And automatic pre- processing of information is removing the noise from selected information.
3. Generalization- extracting patterns for the user query he/she made and finding information.
4. Analysis- analyze the patterns and verify and validate them

below given figure showing all steps involved in crawling process. a crawler gives the required links. the information these links has always according to the user query.

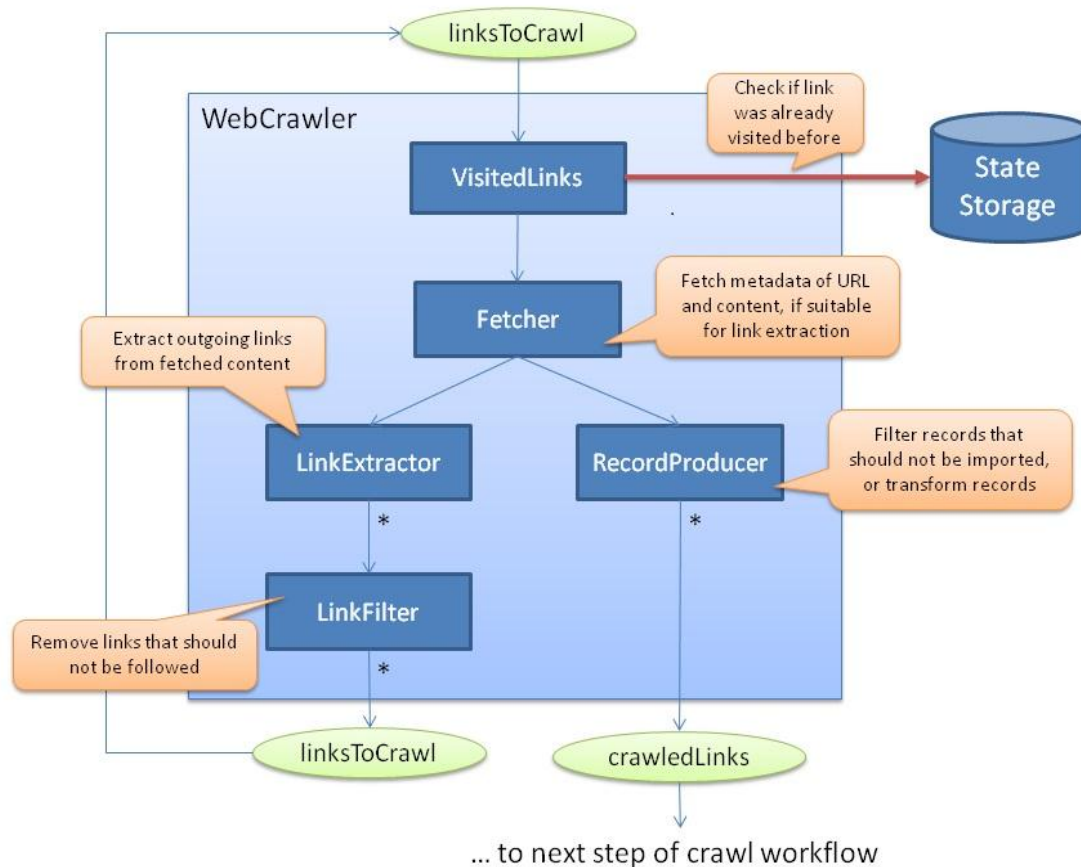


Fig.1. Web Crawling Architecture

1.2 Search Engine (SE)

Basically a search engine is software tool to retrieve information from web resources. These resources can be globally distributed and can be heterogeneous in nature.

Access to heterogeneous distributed information.

- Heterogeneous in creation
- Heterogeneous in thought processes
- Heterogeneous in exactness

A source of unending information.

Examples-

- GOOGLE
- YAHOO
- BING

Working of search engine-

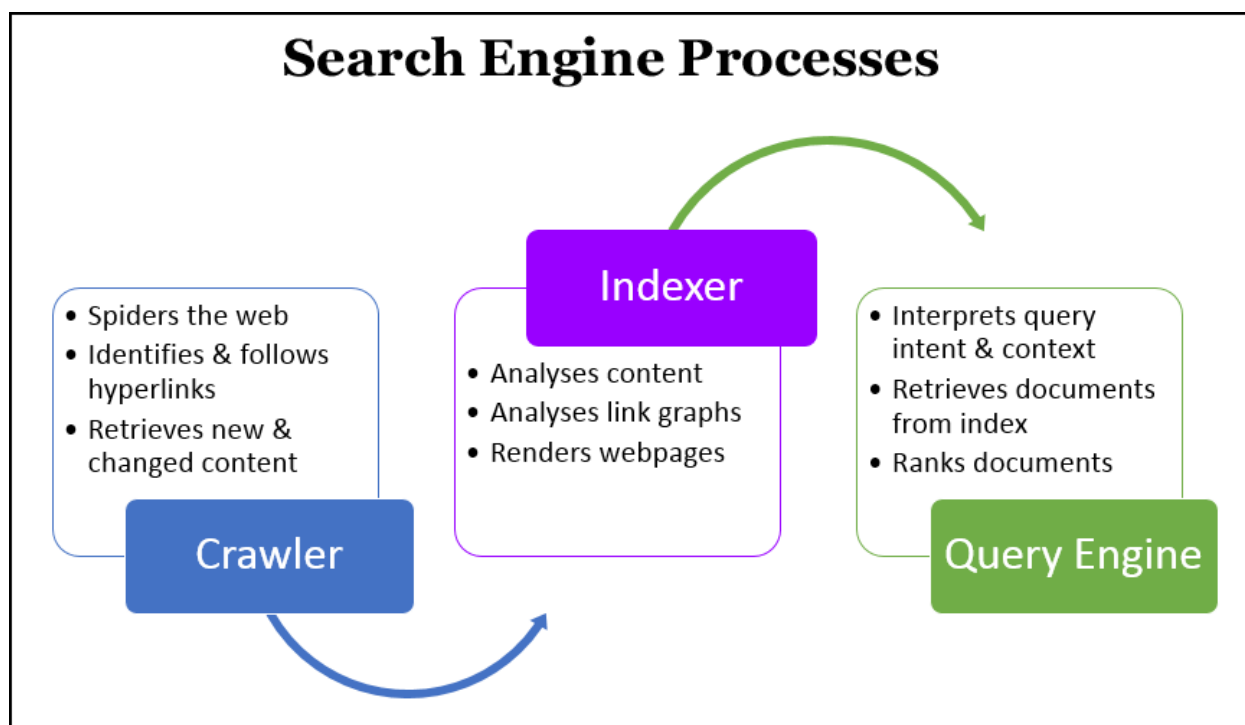


Fig.2. Search Engine Process

- Initially User make a query.
- The web crawler search for the web documents for the query, and take all the search results.
- The websites or the links having related information about the query will be fetched and shown as result on screen by search engine.
- The ranking and the ordering of web documents completely depends on search engine.
- these results based on text matching.
- Indexing of each website done by indexer so that the user can access the information.

WEB DOCUMENTS RANKING ARCHITECTURE

Every search engine use its own technique or algorithm to rank the web documents. Web pages that are highly ranked on many search engines are likely to be more relevant in

providing useful information. Every search engines scores a website and mostly these scores are different for different search engines. Because of using different scoring techniques it's possible that a website which is highly ranked on one search engine is lowly ranked on other search engine. Hence there always a research needed to develop a better technique to rank web documents.

WEB SEARCH RANKING

The web search engine is a software system that is designed to retrieve information on World Wide Web. The search engine results may be of any type it could be an image, web pages and any type file or collection of these types. the ranking of web pages in a search engine means the position at which a particular website will appear in the result of user query.

FACTORS AFFECT THE RANKING

- many factors affects the ranking including-
 1. age of site
 2. quality of site's link portfolio
 3. relevancy of page
 4. level of competition
- according to a survey google using 200 factors to rank the web documents and this ranking can't be controlled by the owner of website.

1.3 Page Rank

Page Rank actually represent the importance of a web page or a website. The page rank is numeric value, page rank is a google technique to measure the "**importance**". Page rank place the more important pages up in search engine result of a user query when it is displayed. When a page links to another page it is said that page effectively cast vote for other page. It calculates the page's importance from the votes cast for it. Each vote matters a lot because it is used to rank the search results.

- First the searching done based on keywords
- Ranking done of the results before display

Page Rank takes back links into count and propagate the ranks through the links. A page has high rank if the sum of its back links is high.

Page rank given by this equation-

$$P_R = (1-x) + x(P_R(T1)/C(T1) + P_R(T2)/C(T2) + \dots + P_R(Tn)/C(Tn))$$

Where-

P_R – is the page rank of page p

$P(T_i)$ - is the page rank of page T_i

$C(T_i)$ - is out going links from T_i

And x is damping factor where $0 < x < 1$

PAGE RANKING ALGORITHMS (PRA)

Proposing new algorithms for ranking web documents involves learning and using machine learning techniques such as Classification, clustering and regression methods.

Content based ranking (CBR) –

In content based ranking we calculate the relevancy of each link. For a user query, result is produced by search engines. Every individual link of result analyzed and pre-processing of data to done to find the noise free data. The user query is also pre-processed to find the root words for given keywords. Form these root words and their synonyms we build a dictionary. Matching to be performed between the Dictionary for the each data word. If a match found a weight is assigned to the keyword that's how we calculate the rank of web documents here.

Usage based ranking (UBR)-

Usage based ranking algorithm are not so accurate but simple. each time we record the visits for a web query. so here the selection frequency we record for a web pages and rank the document based on that score. but it takes a lot to do like saving pages, making bookmarks, analyzing user query etc

Link based ranking (LBR)-

Link based algorithm do not focus on relevancy or keyword count, web document score rather it count the link between the pages. this is a offline algorithms and mostly static in

nature. The ranking algorithm rank the web pages based on the number of the link between the pages and before a query made it produces result based on the count of inter link of pages.

1.4. Web Mining and It's Architecture

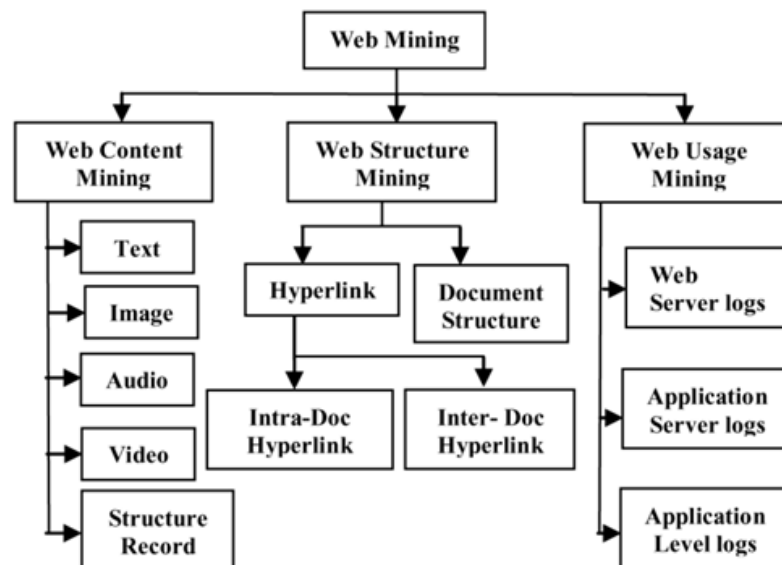


Fig.3. Architecture of Web Mining

Web mining use the traditional methodology to integrate the information or extraction of knowledge over the world wide web. Extraction of data from rich sources makes easy through the web mining process.

1. 5. MOTIVATION

Different search engines using different techniques and organizing the results according to the algorithms they are implemented on. This also affects the user interest to looking for engines to get the desired result for their query. So A novel approach need to be develop to rank the web contents and documents. So a new approach is developed based on keyword and content rather than keyword and page ranking the search engines providing. Based on user query the search engine produced the web contents and results retrieved. Here every result is individually analyzed based keywords and content.

The user query pre-processed to get the root words. These root words are used to dictionary creation and every keyword and page results are matched against this dictionary. If a match

found then a weight is awarded to the word. Finally the total relevancy of each link is computed based on user query by summarizing the total weights whose match found for each keyword and content word.

Similarity and distance measuring we perform in real life. here in machine learning we are doing the same to retrieve the information from the web and ranking or to classify the web documents. we are using k-nearest neighbors classification technique because it is based on measuring distance metric to find neighbors and also using k-means clustering technique because it is also calculate the distance and calculate the relevance of a web page. After all who does not want relevant information in a huge amount and this gives us a positive.

CHAPTER 2: LITERATURE REVIEW

A concealed database alludes to a dataset that an association [4] makes available on the web by enabling clients to issue questions through an inquiry interface. As it were, information obtaining from such a source isn't by following static hyper-joins. Rather, information are acquired by questioning the interface, and perusing the outcome page powerfully created. This paper [7] cures the issue by offering calculations to remove all the tuples from a concealed database. Our calculations are most likely proficient, specifically they achieve the errand by performing just few questions, even in the most pessimistic scenario. We additionally set up hypothetical outcomes showing that these calculations are asymptotically ideal – i.e., it is difficult to enhance their proficiency by in excess of a steady factor. The inference of our upper and lower bound outcomes uncovers noteworthy knowledge into the attributes of the fundamental issue. Broad tests affirm the proposed systems work exceptionally well on all the genuine datasets inspected.

There is an extraordinary measure of profitable data on the web that can't be gotten to by ordinary crawler motors. This part of the web is normally known as the Deep Web or the Hidden Web. We depict a model shrouded web crawler [3] ready to access such substance. Our approach depends on furnishing the crawler with an arrangement of area definitions, every one depicting a particular information gathering errand. The crawler utilizes these depictions to recognize pertinent inquiry shapes and to figure out how to execute inquiries on them.

The customary hunt engine [27] work to creep, file and inquiry the "Surface web". The Hidden Web is the eager on the Web that is open through a html shape not general web crawlers. This Hidden web frames 96% of the aggregate web .The concealed web convey the top notch information and has a wide scope. So concealed web has dependably stand like a brilliant egg according to the specialist. The records re established by a shrouded web crawler are more legitimate, as these archives are open just through powerfully produced pages, conveyed because of a question.

In a developing universe of web, a great many website pages are included day by day. Be that as it may, just approx. 0.03 percent part of pages are recovered by all the web crawlers. Rest of the website pages or archives are profound web assets. We propose a double layer structure [19], to be specific Smart Web Crawler, for proficiently recovering profound web interfaces. Shrewd crawler comprises of two layers: Site-disclosure and inside and out

Crawling. Webpage disclosure layer finds the sparsely found profound sites from given known parent destinations utilizing Reverse Searching and centered slithering.

Meta search [8] defined as searching for a query at multiple system s simultaneously. Here what we do is we send the user query to multiple search engines simultaneously and get the result from multiple search engines and merge the result and ranking done on all results after that the results are presented to the user.

For each search engine we have to do modifications for its scoring techniques and ranking algorithms [16]. These decisions made by user based on the query keywords not all the information he/she wants. So the main thing here is the ranking aggregation techniques. So merging of results playing a key role here and the efficiency totally depends on the merging algorithms. Here we are investigating different merging techniques there scoring approaches and comparing them against each other.

A productive utilization of substance construct positioning with respect to positioning [11] sites was recommended by Zhu et al. In this approach the arrangement of hyperlinks are separated into perusing joins, proposal joins. A chart of blog webpage is made by expansion of shrouded connect in light of the substance investigation. An idea of certain and unequivocal positioning was presented in this method.

A strategy proposed by Fey Zania et al to rank the reports by developing an associated chart of semantic records. This chart considers just the semantic connections between the records [15] .Thus a novel approach of substance based positioning was advanced as needs be by extricating the certain positions. Distinctive weights can be utilized to recognize the diverse semantic connections.

A compelling way to deal with utilization based positioning [4] utilizing metaphysics was advanced by Jun tooth et al . The weight computation was finished utilizing cosmology tree. This weight figuring assesses use data, examples and structure. The whole cosmology tree was changed over into a weighted diagram and afterward avocation was connected to compute the last weight which was then utilized for seeking, positioning and strife illuminating.

A vital survey[9] on page positioning calculation was exhibited by Selvan .The whole overview gave a proficient correlations on these calculations .The primary spotlight was on the crucial page positioning calculations like hits and centered rank .The examinations were drawn on the parameters of benefits, negative marks, execution and significance.

An effective approach of web service recommendation based on usage based ranking[14] and QOS preferences was presented by Kang .This algorithm is effective in classification and extraction of the results.

In a novel approach in which crawler downloads just applicable records exploiting vagrants and in this manner diminishing the heap on server it downloads just those site pages that are significant to a specific subject or an arrangement of points and gives data in a customized see thinking about just client inclinations.

In Patil has called attention to that the induction and examination of inquiry objectives can have a ton of points of interest in enhancing web index significance and client encounter by joining web use and web content mining [1]. It introduces a weighted procedure to mine the web content taking into account the client needs.

Some positioning calculations [16] have incorporated the investigation of inquiry logs. Navigate information is additionally utilized via web indexes to assess the nature of changes to the positioning calculation by following them. Coordinate Hit 1 utilized past session logs of an offered question to figure positions in light of the Popularity (number of snaps) of every URL that shows up in the appropriate responses of the inquiry. This approach works for inquiries that are oftentimes detailed by clients, in light of the fact that less normal questions don't have enough snaps to enable critical positioning scores to be computed. For less normal questions, the immediate hit rating gives a little advantage.

Zhang and Dong (2002) [11] propose the Matrix Analysis on Search Engine Log (MASEL) calculation, which utilizes web crawler logs to enhance positioning. Snaps are viewed as positive Suggestions for pages. The fundamental thought is to separate from the logs connections of clients, inquiries, and pages. These connections are utilized to evaluate the nature of answers in light of the nature of related clients and inquiries. The approach depends on the recognizable proof of clients of a web crawler, an undertaking hard to accomplish by and by.

There is additionally ongoing related work on question grouping, a few methodologies likewise think about information in inquiry logs. We nand partners (2001) propose bunching comparable questions to prescribe URLs to much of the time solicited inquiries from a web crawler.

They use four notions of query distance:

- (1) Based on keywords or phrases of the query
- (2) Based on string matching of keywords
- (3) Based on common clicked URLs

(4) Based on the distance of the clicked documents in some pre-defined hierarchy.

Befferman and Berger (2000) likewise propose an inquiry grouping strategy [26] in light of basic clicked URLs. From an investigation of the log of a well known internet searcher, Jensen and associates (1998) inferred that most questions are short (around 2 terms for every inquiry) and loose, and the normal number of pages clicked per answer is low (around 2 ticks for every question). In this manner, ideas (1)– (3) are hard to manage by and by, on the grounds that separation frameworks between questions created by them are exceptionally scanty. Thought (4) needs an idea scientific classification and required the clicked records to be characterized into the scientific classification also.

CHAPTER 3: PROPOSED WORK

3.1 Problem Statement

In today's world, the data reside online and information retrieval becoming easy every day. Search engines are the tool we use regularly to get the required result. These tools are very efficient in terms of performance and accuracy of fetched documents. But for a particular query these tools may have a huge data records, which comes in the form of web documents in users hand. For making the working of search engine more efficient when it comes to the ranking of web documents. Our work shows that the web page ranking gives more accuracy when we use clustering in content based search engine. Web content mining and information retrieval is an ample and powerful research area in which retrieval of relevant information from the web resources in a faster and better manner. Web content mining improves the searching process and provides relevant information by eliminating the redundant and irrelevant contents. here unlike the page ranking algorithms we are using weighted (content) based ranking algorithm and comparing the results .further we are also using clustering technique and classification for better results .results shows this way the ranking of web documents or pages gives better performance in terms of precision recall and f-measure.

3.2 Related Work

To rank web documents based on relevancy factor. User makes a query and framework or search engine searches all the available web pages related to the text. Here user gets the information with a little effort and we can get or retrieve information from structured and unstructured documents or information resources. Along with this we also use some methods to remove unwanted and unnecessary data content. Most of the people now a day's rely on search engines to get information but when a user uses search engine like google, yahoo and bing, they also return enormous quantity of data having relevant and irrelevant information both. So getting relevant information from this much amount of data become interesting to the research area.

- Words are fully matched against the dictionary.
- User query made to get results and these results are pre-processed.
- Pre-processing step is important in text content mining.

- Real data can be incomplete, noisy and irrelevant to get the quality in data pre-processing done.
- Knowledge discovery becomes easy because we need quality data in order to make quality decisions.
- Keywords and content words are pre-processed to remove noisy data.
- Dictionary made after the pre-processing step from user query.
- The keywords and content words are match against the dictionary and if a match found a point is given else no point will be awarded
- Weighted technique work on above principles.
- After all these steps content words and keywords summarized and normalized so that the cumulative total will be less than or equal to 1.

ALGORITHM

This Algorithm is based on relevancy and weighted approach. The algorithm takes extracted web data as input and gives us ordered data or web documents links as output.(**for each link**).

- Extract search results for a user query say for a where $1 < a < N$ and results are S_a
- Get the root words for the query using pre-processing. Say it R_b where b is $1 < b < N$
- Make the dictionary from root words.
- Get the total keyword from search result S_a and let's say it's A .
- Get the total content words from search result S_a and let's say it's B .
- Compute strength of keywords (SK)

$$SK=1/A$$

- Compute strength of content words (SC)
- Compute $\sum(SK)$ and $\sum(SC)$

$$SC=1/B$$

Let's say we have a weighted variable W where $0 < W < 1$

- Then **relevancy(R)** will be
- $R= W*\sum(SK) + (1-W)*\sum(SC)$
- **The link with higher Relevancy have higher rank.**

Performance evaluation done based **precision, recall and accuracy.**

Let's say

We have test data and actual data then

Precision (P) = true positive/(true positive + false positive)

Recall (R) = true positive/(true positive + false negative)

Accuracy = (true positive + true negative)/(true positive + true negative + false positive + false negative)

F-measure = $2 * P * R / P + R$

Let's say for a query we get these results

- Manual ranking done by users.
- One row for ranking done on proposed approach
- One row ranking done by some random search engine
- After comparing them we get the following results-

Documents	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
S.E.Rank	1	2	3	4	5	6	7	8	9	10
Manual	5	9	4	2	1	3	8	6	7	10
Given approach	5	3	4	2	1	8	9	6	7	10

- Hence from the given data we have

$$P = .7$$

$$R = .88$$

$$\text{Accuracy} = .9$$

$$\text{F-measure} = .77$$

which is quite high, the clustering (k- means) technique and classification technique (k-nearest neighbour) technique making this area even more interesting to reduce the noise and producing more accurate result through a efficient web page ranking.

3.3 Used Techniques

Web page ranking depends on various factors in content based web ranking each result is different .different search engines display different result. By far we have implemented and used many algorithms of page ranking they are giving better outputs. Proposed work focusing on Clustering and Classification technique, Aim is to getting good information retrieval and a relevant ranked web document. During the study i found that the search engines will be more precise if we apply clustering on URLs and creation of clusters will give good results.

ALGORITHMS

K-MEANS

K-NEAREST NEIGHBOUR

K-MEANS CLUSTERING

During the research work we have analyse the data sets and calculate the respective results. These results are found different for different data sets, For each we have plotted precision graph and dataset to time graph. These graph shows that the precision of existing approach and the proposed solution are different and proposed solution rank the web pages in efficient manner with higher precision.

K- Means clustering is a unsupervised learning technique.

K- Means Algorithm takes the un grouped and unlabeled data and group them into clusters. the clusters to be done depends on the variable K .here the data is marked and grouped into clusters on the basis of similarity and distance matrix.

The results of algorithm:

- The Centroid of clusters used to label new data.
- Assigning labels for the training data, here the data set is assign to a particular cluster.

Before looking to a cluster, clustering allows us the study of Centroid of each cluster which shows the properties and weights of group members. Every Centroid of a bunch is a gathering of highlight esteems which characterize the subsequent gatherings. Analyzing the Centroid highlight weights can be utilized to subjectively decipher what sort of gathering each group speaks to once the gathering done its easy to appoint new information to its bunch.

K-means calculation is most normal and famous grouping technique that is broadly utilized as a part of numerous applications and it falls under the apportioning calculations that points in developing the different examples and assesses them by utilizing some standard. With the given accumulation of n information, k distinctive groups are shaped with each bunch having an one of a kind centroid (mean) and in this manner the dividing is made. The letter k depicts the quantity of bunches should have been made. At the point when number of n objects is to be gathered into k bunches, K group focus is to be introduced. Each question will be given to the nearest bunch focuses and. the focal point of bunch is refreshed each time until the point that condition of no change happens in the each group. The components in each group will be in close contact with centroid of that specific bunch and will be diverse to the components having a place with different clusters.

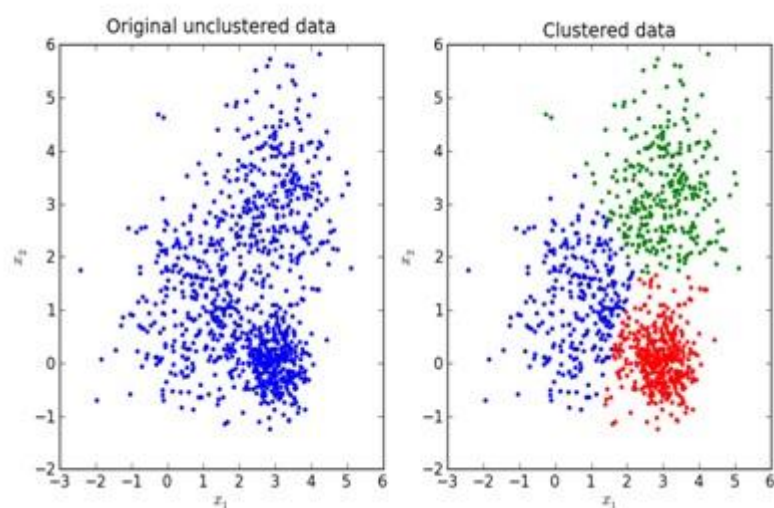


Fig.4 Unordered Data VS Clustered Data Comparision

The whole of the disparities between the point and centroid communicated by particular separation is utilized as the goal work. Add up to intra-bunch fluctuation portrays the aggregate of the squares of the blunder between the point and separate centroids. Information is isolated in various bunches, which are generally been sufficiently far separated from each other spatially, in Eucledian Distance, to have the capacity to create successful information mining comes about.

Each bunch has a middle, called the centroid, and an information point is grouped into a specific bunch in view of how shut the highlights are to the centroid.

This Algorithm works iteratively, and in each progression it lessens the separation between the information focuses and their Centroid. This how it gives ideal answer for the given datasets. The calculation inputs are the quantity of bunches K and the informational

collection. The informational collection is an accumulation of highlights for every datum point. The calculations begins with beginning assessments for the K centroids, which can either be haphazardly produced or arbitrarily chosen from the informational index.

This algorithms works mainly in two steps:

ASSIGNMENT OF DATA –

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance.

UPDATION OF CENTROID-

The centroid of each cluster is the mean value of the data sets falls in a particular cluster.

hence each time when a data is added to a cluster, centroid has to be re calculated, this value set to be new centroid of cluster.

The algorithm keep on working between these two steps and don't stops untill it finds one of the following conditions:

1. The value of the centroid coming same means mean stop changing.
2. The sum of distance of any cluster become minimum.
3. May be a maximum numbers of iteration occurs.

The more number of clusters represent minimum centroid to data point distance the valuse of K may be randomly chosen or can be generated. Various different systems exist for approving K, including cross-approval, data criteria, the data theoretic hop strategy, the outline technique, and the G-implies calculation. Furthermore, checking the dispersion of information focuses crosswise over gatherings gives understanding into how the calculation is part the information for every K.

Work is as follows: first decide the value of K THEN

- Place K focuses into the space spoke to by the items that are being bunched. These focuses speak to starting gathering centroids.
- Assign each protest the gathering that has the nearest centroid.
- When the sum total of what objects have been appointed, recalculate the places of the K centroids.
- Repeat Steps 2 and 3 until the centroids never again move. This delivers a detachment of the items into bunches from which the metric to be limited can be figured.

The strategy will dependably end, the k-means calculation does not really locate the most ideal setup, relating to the worldwide target work least.

ADVANTAGES OF K-MEANS

1. Produces the scalable solutions.
2. simple and easy and used for undirected knowledge discovery from a large data set.
3. this clustering technique has many application like image processing, pattern recognition and neural networks.
4. gives best results when the dataset is large and distinct and having well separable distance.

K-NEAREST NEIGHBOUR

The k-Nearest Neighbors calculation (or k-NN for short) is a non-parametric technique utilized for characterization and relapse. In the two cases, the info comprises of the k nearest preparing cases in the element space. The yield relies upon whether k-NN is utilized for arrangement or relapse. In k-NN characterization, the yield is a class enrollment. A question is grouped by a greater part vote of its neighbors, with the protest being doled out to the class most regular among its k closest neighbors (k is a positive whole number, normally little). In the event that $k = 1$, at that point the question is just relegated to the class of that solitary closest neighbour.

In k-NN relapse, the yield is the property estimation for the question. This esteem is the normal of the estimations of its k closest neighbors. k-NN is a sort of occurrence based learning, or languid realizing, where the capacity is just approximated locally and all calculation is conceded until characterization. The k-NN calculation is among the easiest of all machine learning calculations.

Both for arrangement and relapse, it can be helpful to appoint weight to the commitments of the neighbors, so that the closer neighbors contribute more to the normal than the more far off ones. For instance, a typical weighting plan comprises in giving each neighbor a weight of $1/d$, where d is the separation to the neighbor.

Distance functions

Euclidean	$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^k x_i - y_i $
Minkowski	$\left(\sum_{i=1}^k (x_i - y_i)^q \right)^{1/q}$

Fig.5 Distance Functions

The neighbors are taken from an arrangement of items for which the class (for k-NN characterization) or the question property estimation (for k-NN relapse) is known. This can be thought of as the preparation set for the calculation, however no unequivocal preparing step is required. The preparation cases are vectors in a multidimensional element space, each with a class mark. The preparation period of the calculation comprises just of putting away the component vectors and class marks of the preparation tests. In the order stage, k is a client characterized consistent, and an unlabeled vector (an inquiry or test point) is arranged by allotting the name which is most incessant among the k preparing tests closest to that question point.

A usually utilized separation metric for constant factors is Euclidean separation. For discrete factors, for example, for content grouping, another metric can be utilized, for example, the cover metric (or Hamming distance).

The primary favorable circumstances of KNN for characterization are:

- Very straightforward execution.
- Robust as to the pursuit space; for example, classes don't need to be directly divisible.
- Classifier can be refreshed online at almost no cost as new occasions with known classes are displayed.

- Few parameters to tune: separate metric and k.

A shortcoming of the k-NN estimation is that it is fragile to the area structure of the data. The figuring has nothing to do with and isn't to be mixed up for k-suggests, another popular machine learning technique.

3.4 Proposed Solution Architecture

Search Engines play a vital role in searching for a user query and display the relevant results to the users. But for a very small query a search engine produces a huge data as results. But this data may not be noise free. Noise may consist any unwanted data/text. When a query is made by a user a search engine search for the related data and fetches all the results. But for a better information retrieval data has to be noise free and consistent. Studies also shows that if we summarize the fetched data in a sophisticated manner the output of search engines gives more precision.

Various search engines rank their web pages through various ranking algorithms and shows different outputs. Ranking of documents is a crucial task, and using clustering with content based ranking makes this task even more accurate and relevant.

ARCHITECTURE OF PROPOSED SOLUTION

Ranking of web documents is a hot topic to research on, till now many techniques and algorithms proposed to rank the web documents. This takes my attention to this area, because no user will ever want inconsistent and unorganized information for his/her query. The time is also a factor to motivate the researchers in this particular area. This work is also efficient in terms of time when it is compared to the basic content based ranking of web documents.

The fetching of data or web pages and arranging them in a efficient way starts from the very initial phase (making query to search engine) and ends with a complete and relevant ranked web document to the user.

Previously a lot of work has already been done in this field, thus this area keep on motivating us to move further and getting better and better results through new proposed solutions and techniques. This explains the whole structure of previous and proposed solutions and all the important stages through which this work went by, getting required results and that too in efficient time is tough task.

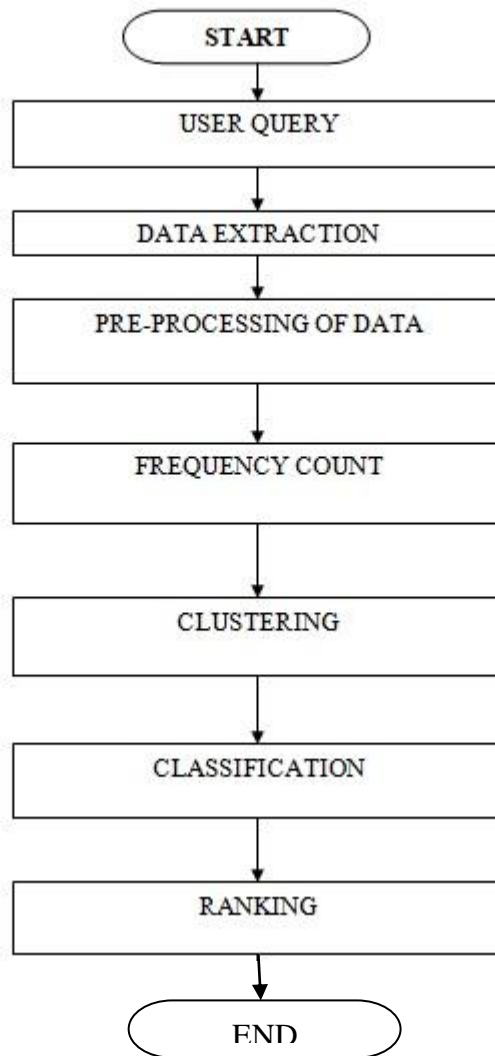


Fig. 6 Architecture of Proposed Solution

In this era of internet, search engines are the rich sources of information having vast amount of relevant and irrelevant data. We as a user always want relevant data. This task is done by the search engines automatically they use many web page ranking algorithms to put their results in a efficient manner.

Above block diagram shows each and every step this work went through. After these several basic operations or steps the search engines produces optimized web documents in particular order having relevant data for users information retrieval.

STEPS

USER QUERY-

User can pass any information to the search engines according to his/her needs, this information may be a set of keywords representing what a user wants. Search engines works

on the query and found the desired data. This fetched data initially comes as result to the user these web pages are already been ranked using some ranking algorithm. Like every search engine our work also focus on keeping the most important pages on the top.

DATA EXTRACTION-

When a user make a query, search engine looks for the relevant data from the online resources. And fetched all the data related to the query and represent them in the form of links or urls. Then these urls are ranked by the search engine.

Search engines perform data extraction through web crawling, the crawling process is nothing but the process of fetching meaningful data from all the available resources. And after this all the data been saved. Data is really very large hence we have to mine the meaningful information from the raw data searched by a search engine.

DATA PRE-PROCESSING-

data pre-processing is a must task to perform, data pre-processing is nothing but the process of removing the noise from the extracted data crawled by the search engine. Initially the data fetched by search engines contain words or text like(., : , ; , & , # , @ , - , _ , and other irrelevant text or symbols etc). This data has to be removed from the results. After the pre-processing phase of extracted data we get the relevant and important data.

Removal of these unwanted keywords and symbols makes the data suitable for next phases. Hence importance of this step in web document ranking is quite high.

FREQUENCY COUNT-

The proposed method and the rest of existing techniques in content based search engines, we perform keyword count and make a dictionary. The frequency count represent the presence of each keyword or word. Here as an input we gives all the pre-processed data to count the frequency. After making a dictionary from the user query keywords and their synonyms, one more dictionary is to be made based on the frequency count of each word or keyword in data. The query dictionary keywords then matched against the dictionary of the found processed data. For each match we assign a weight to the keyword called relevancy weight. This is how we calculate the weight of every link or URL, then based on that relevancy factor search engines rank the web documents/pages.

CLUSTERING-

Proposed work uses K-means algorithms as clustering algorithms, here the number of clusters are variable and we can change the clusters according to the need. Performing clustering on processed data produce the clusters, the URLs get divided into clusters according to the their relevancy factor and similarity to one another. Web page falling into same clusters shows similarity in terms of Relevancy, Size of data in link, Word count, and keyword matching.

Pseudo code

1. Initialize **cluster centroids** $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

For every i , set

$$c^{(i)} := \arg \min_j \|x^{(i)} - \mu_j\|^2.$$

For each j , set

$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

}

The first line says to choose centroid of clusters randomly, where initially the numbers of clusters i.e the value of K is unknown and hence chosen wisely and in starting of algorithms. After this In each Iteration the value of Centroid changes during the clustering of input data the algorithm stops when the number of iterations get over or the value of centroid remain same. One more case is also can be terminating condition for k-means which is when the distance between the test data and centroid becomes very small. For each set i , we are calculating the data items fall into the same categories or clusters. And for each set j , we are updating the centroid for each cluster.

CLASSIFIATION-

When the clustering done on the searched result, the Links for respective resource of data, classified into clusters and the whole cluster have some relevancy in terms of the content matched against the required query made by the user. The classification technique is such a brilliant technique to rank web documents. Here the produced results of clustering, further passed to k-n neighbour algorithm to perform the classifications of pages using this algorithm based on distance measuring matric. This Algorithm calculate the distance for each dataset to

the nearest neighbour from the training classes. The data set showing maximum similarity and minimum euclidian distance classified to the respective class, this is how the more relevant links are classified into same class. This classification produce more precision when compared to the traditional way of page ranking.

RANKING-

Finally the web documents are ranked, and outputs are displayed on the screen, the time complexity and precision comes as outputs. Work shows more relevancy in links and web pages ranked on the basis of relevancy and content count gives more precise web documents. When for a user query search engine produce more relevant and accurate ranked documents, the task of information retrieval look more easy and efficient.

3.5 Working

Above architecture clearly shows the main phases of the studies. During the research and literature survey we have seen all ranking algorithms and some other techniques to mke the process efficient.here when a user make a query it goes to server and user waits for server response. At server side all these operations takes place. When query response given as web documents we perform content based maching on each links data, the keyword dictionary we make for keywords and their synonyms match against the words of pre-processed data. The basic alorithms in content based ranking counting the word count for each keyword and calculating the relevancy of web pages, through assigning weights accordingly. In proposed solution we count the words forms the keyword dictionary.

Instead of just counting the strength of the keyword we are also using the total words in data and see howmuch match occur from how much data.

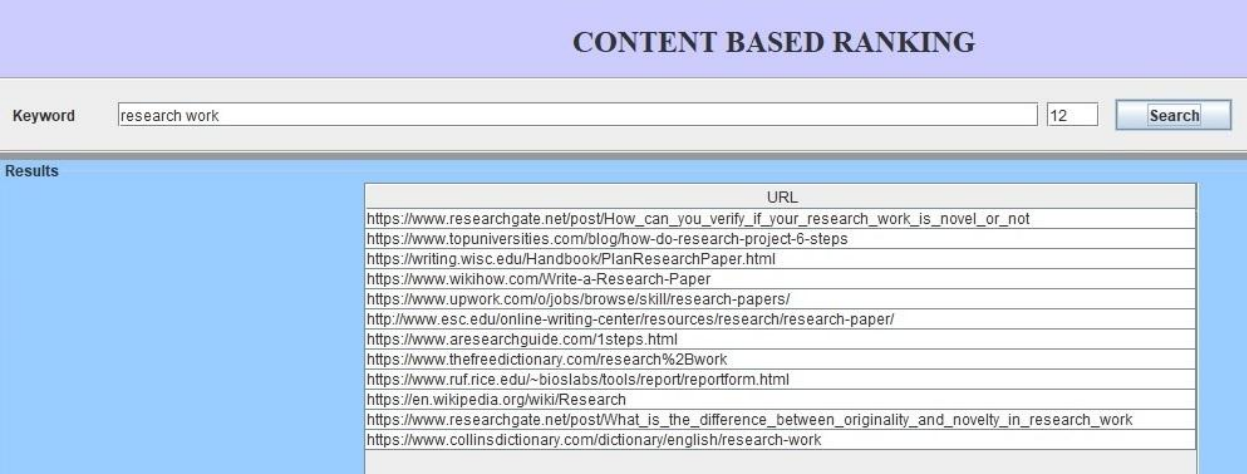
A ratio is calculated for **(matched keyword count/total count of words)** , after calculating the values for each links or web documents the data is pass to the clustering algorithm where the high relevant links grouped into the clusters. A good classofication technique here classify web documents to a particular cluster, the k-value represent the relevancy of web documents in our proposed approach. Calculated k values are used to rank the web pages/documents. Project also **covering the respective data size to time comparision and performance compaerision of existing and proposed technique**. And we found positive results.

CHAPTER 4: SIMULATION AND RESULTS

Ranking Web documents is a necessary task has to be performed by the search engines for a particular query. Experiment shows this task becomes even more precise when we use clustering and classification techniques to rank the pages. Traditional ranking algorithm rank the web pages but that is not so efficient in terms of precision, recall and also data we get in response from the server is need to be pre-processed and consistent. The work proposed here is taking data from online resources hence is dynamic in nature. We get different results for different data inputs. For a given query we may have huge data as output. The searched links gives us required data as outpost so ranking of those links become a crucial task. When a query is made search engines gives links or web documents in output.

USER QUERY RESULTS

Figure shows the output links for a user query.



URL
https://www.researchgate.net/post/How_can_you_verify_if_your_research_work_is_novel_or_not
https://www.topuniversities.com/blog/how-do-research-project-6-steps
https://writing.wisc.edu/Handbook/PlanResearchPaper.html
https://www.wikihow.com/Write-a-Research-Paper
https://www.upwork.com/o/jobs/browse/skill/research-papers/
http://www.esc.edu/online-writing-center/resources/research/research-paper/
https://www.aresearchguide.com/1steps.html
https://www.thefreedictionary.com/research%2Bwork
https://www.ruf.rice.edu/~bioslabs/tools/report/reportform.html
https://en.wikipedia.org/wiki/Research
https://www.researchgate.net/post/What_is_the_difference_between_originality_and_novelty_in_research_work
https://www.collinsdictionary.com/dictionary/english/research-work

Fig.6 USER QUERY RESULTS

The number of Links or URLs coming as output are variable and we can easily set them to random number. By default we have set the value to the 14 to have average no of results as output and getting good comparison graph between the existing and proposed work.

EXTRACTED DATA

The data is extracted based on the user query, data from each link then saved to a depository. This takes a little time because the data is in bulk amount. The extracted data must be noise free and hence pre-processing process came as an important task, after that the data becomes more relevant.

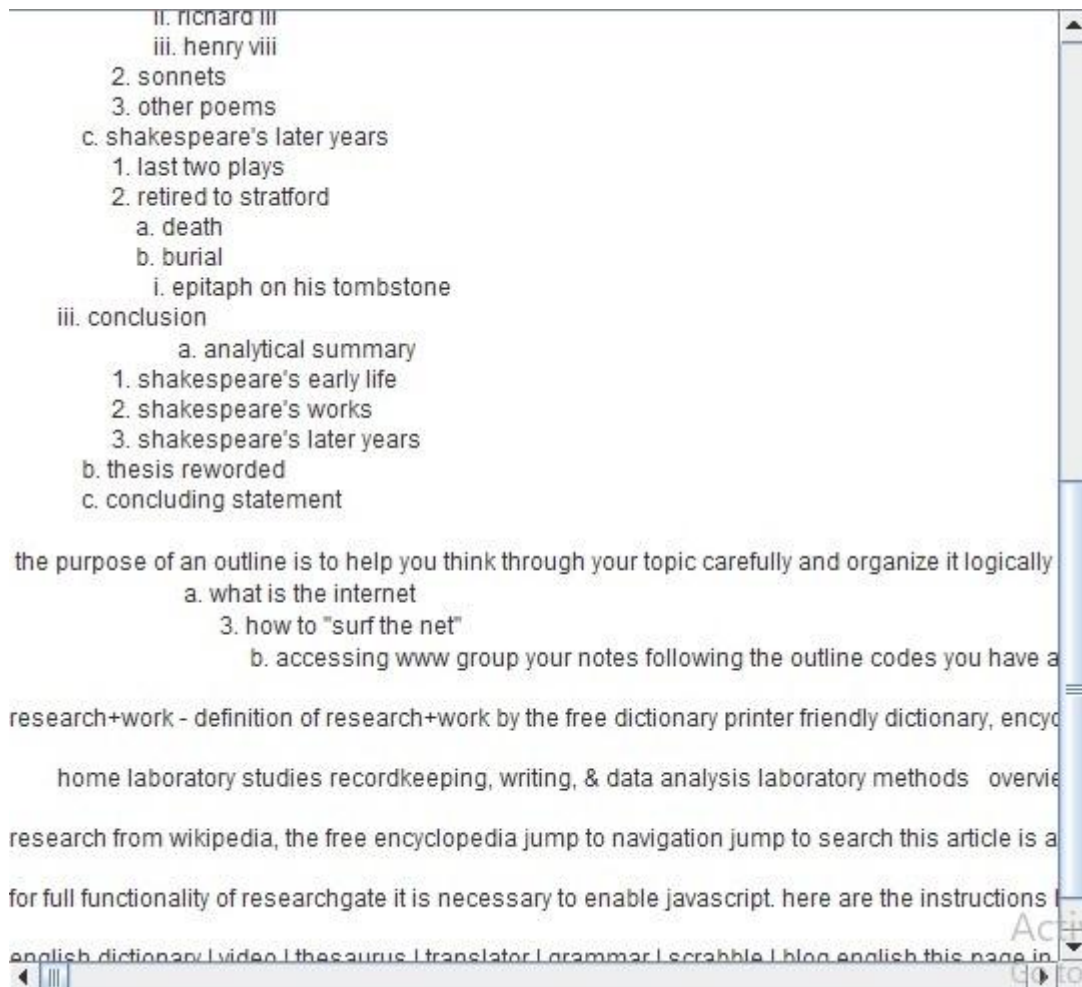


Fig.7 EXTRACTED DATA

Above figure shows a small snap of extracted data for a user query “RESEARCH WORK”, the extracted data contain many unwanted text called noise.

PRE-PROCESSING

This phase makes the extracted data more precise and noise free. We have chosen almost all unnecessary and unwanted words/symbols and text etc. During pre-processing all these unwanted data removed.

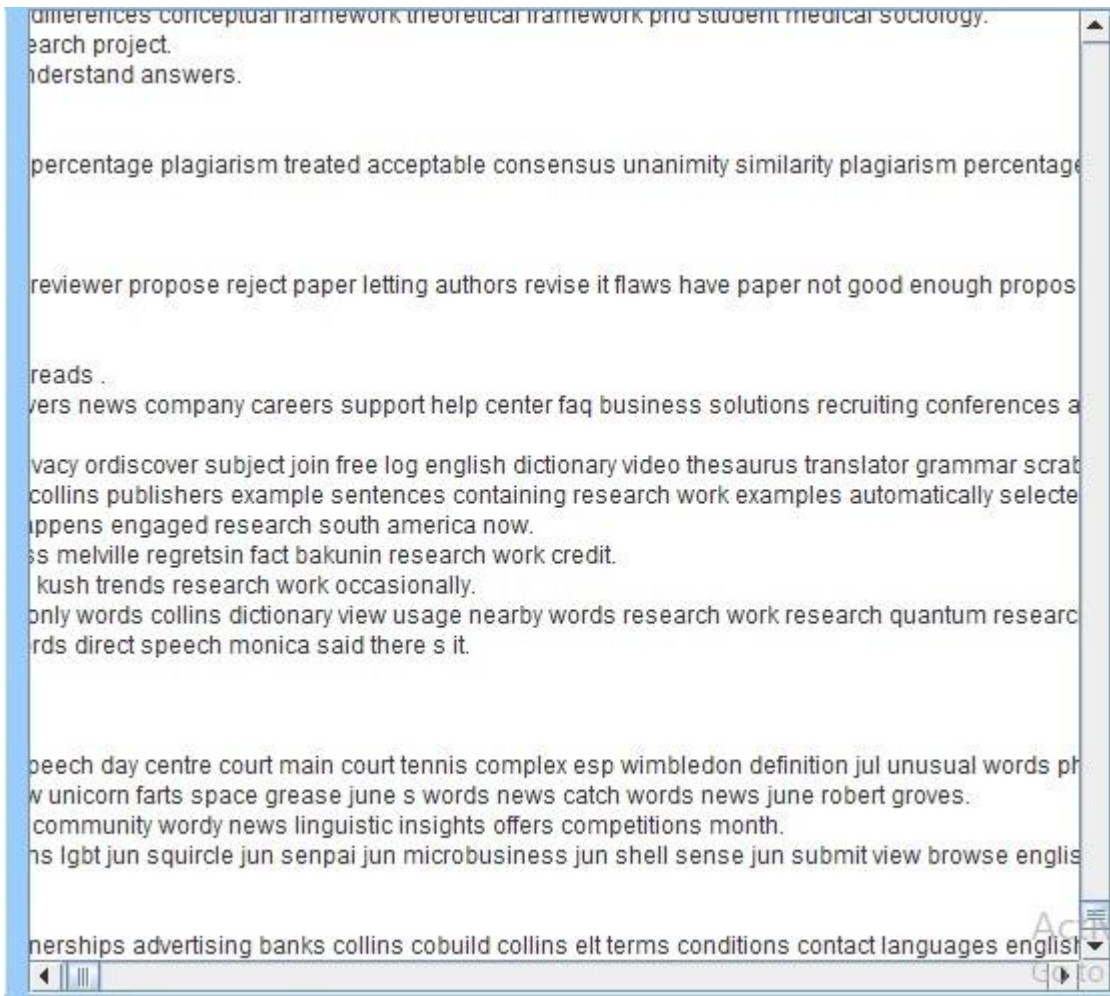


Fig.8 PRE-PROCESSING

Above figure showing the unwanted text/symbol free data extracted by search engines through web crawling. A huge unwanted amount of data been removed in this phase.

FREQUENCY COUNT

as we know in content based search engines, we studied that to make a efficient ranking procedure we need to count the terms or words of the data. This phase count the frequency of text present in extracted data.

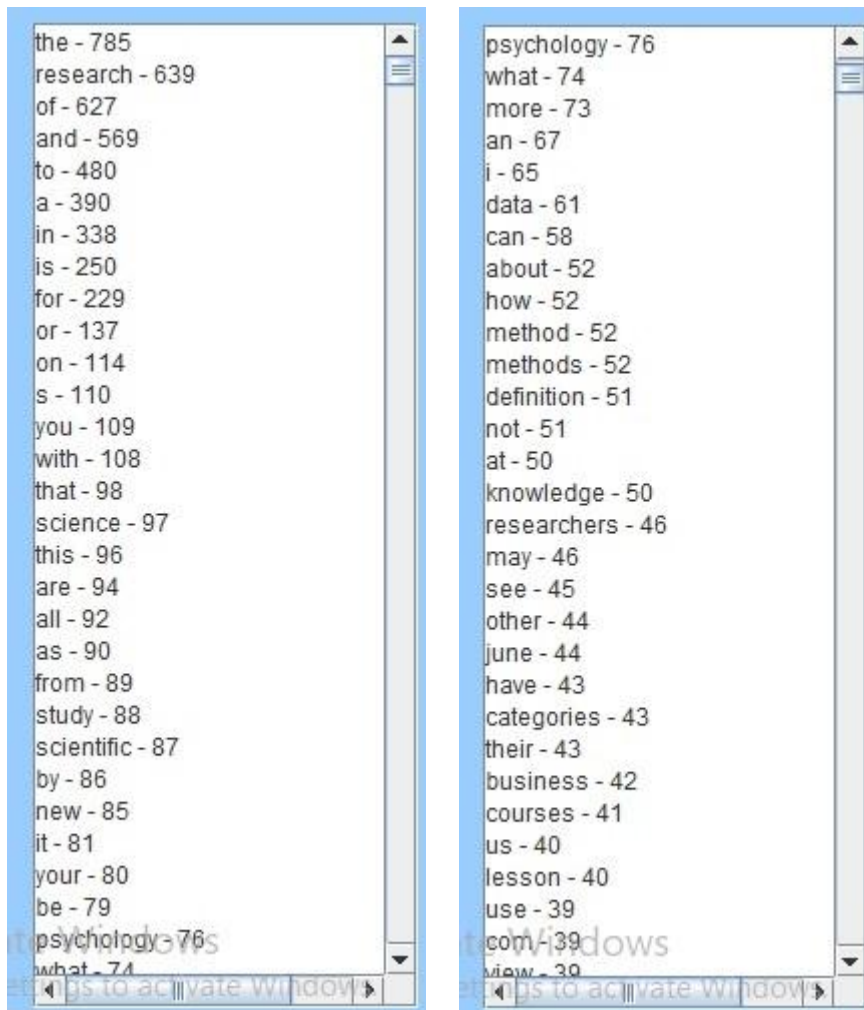


Fig.9 FREQUENCY COUNT

Above figure shows the count or frequency of data keyword present in the pre-processed data.

Making dictionary with these words and their synonyms is used to match against the user query and respective weight is assigned to the keyword. These weights sometime called as relevancy factor or relevancy of any document.

CLUSTERING RESULTS

K-MEANS is the clustering algorithm we have applied on URLs of the fetched data. After clustering we get clusters of URLs based on similarity and relevancy of each document.

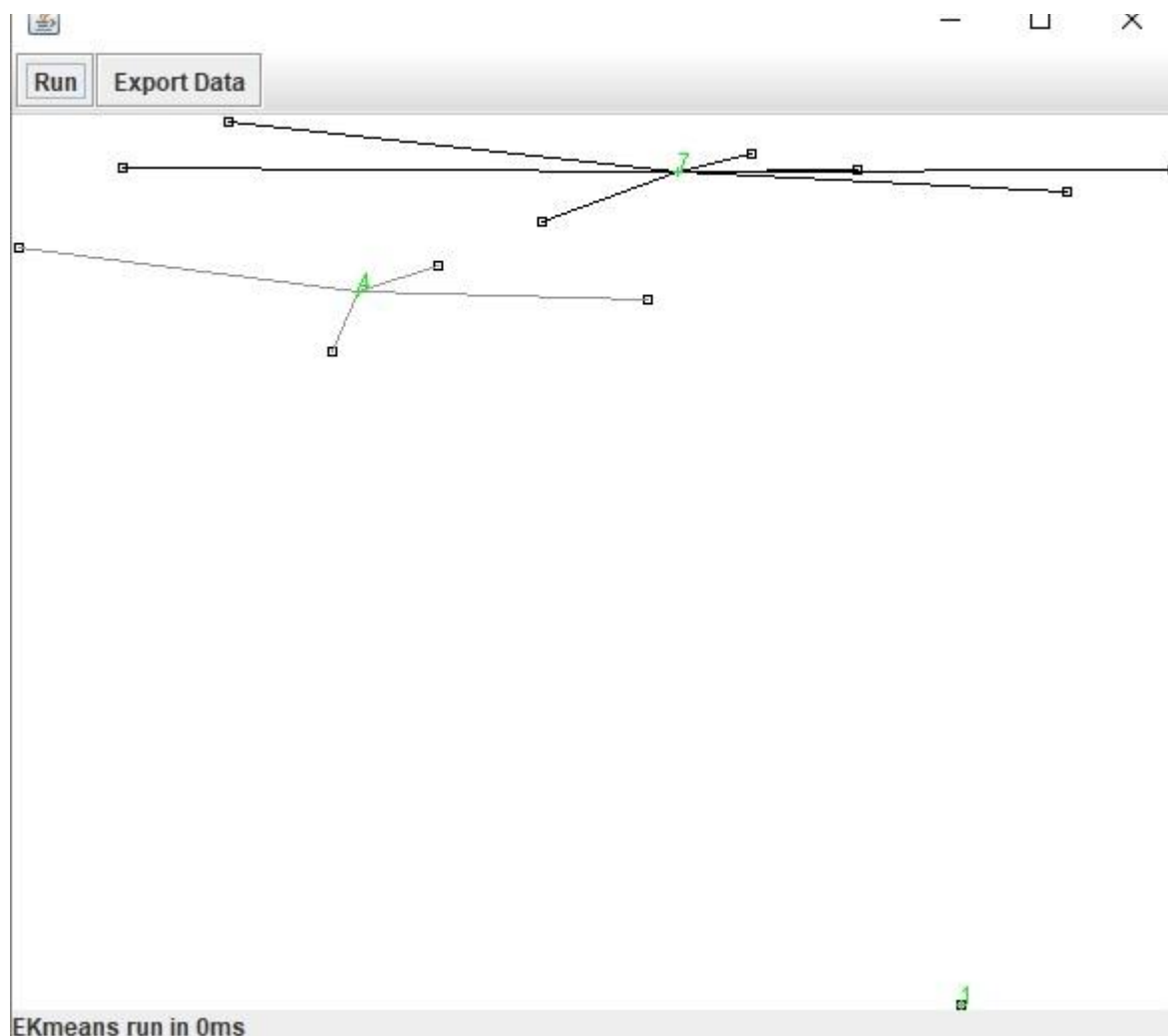


Fig.10 CLUSTERING RESULTS

Above figure is the result of clustering done on 12 URLs searched for user query “research work”. We have variable numbers of clusters in this particular example we have taken only three. But for more précised results we can go for more. More will be the clusters data set divided into more groups and hence search engine results will be more precise.

CLASSIFICATION RESULTS

The work here done is the classification of URLs . all links are classified into clusters and which particular link fall in which cluster based on Euclidian distance.

URL	Cluster
https://www.researchgate.net/post/How_can_y...	Cluster - 2
https://www.topuniversities.com/blog/how-do-r...	Cluster - 3
https://writing.wisc.edu/Handbook/PlanResear...	Cluster - 3
https://www.wikihow.com/Write-a-Research-P...	Cluster - 2
https://www.upwork.com/o/jobs/browse/skill/re...	Cluster - 2
https://www.aresearchguide.com/1steps.html	Cluster - 3
https://www.thefreedictionary.com/research%2...	Cluster - 2
https://www.esc.edu/online-writing-center/reso...	Cluster - 3
https://www.ruf.rice.edu/~bioslabs/tools/report/...	Cluster - 3
https://en.wikipedia.org/wiki/Research	Cluster - 1
https://www.researchgate.net/post/What_is_th...	Cluster - 3
https://www.collinsdictionary.com/dictionary/en...	Cluster - 3

Fig. 11 CLASSIFICATION RESULTS

We can see which particular URL belongs to which cluster. Results shows that the

Cluster1: have only 1

Cluster2: have 4 links of web documents.

Cluster3:have maximum number of links to documents and its 7.

Hence the cluster 3 having 7 links is the most relevant and these URLs must come first. So user can retrieve his/her information in most efficient manner.

RANKING

Ranking of web documents is a major step for web content based search engines. The ranking is performed on the basis of K-VALUE.

URL	K-Value
https://en.wikipedia.org/wiki/Research	3.404
https://www.upwork.com/o/jobs/browse/skill/re...	2.52
https://www.aresearchguide.com/1steps.html	1.796
https://www.topuniversities.com/blog/how-do-r...	1.579
https://www.wikihow.com/Write-a-Research-P...	1.546
https://www.ruf.rice.edu/~bioslabs/tools/report/...	1.377
https://www.researchgate.net/post/How_can_y...	0.956
https://www.researchgate.net/post/What_is_th...	0.515
https://www.thefreedictionary.com/research%2...	0.412
https://writing.wisc.edu/Handbook/PlanResear...	0.359
https://www.esc.edu/online-writing-center/reso...	0.302
https://www.collinsdictionary.com/dictionary/en...	0.246

Fig.12 RANKING

This value represent the relevancy of web page. First ranked web pages will have the highest k value.

The figure clearly shows that the k-value of <https://en.wikipedia.org/wiki/Research> is **highest** and it is calculated 3.404 where as the web document having URL <https://www.collinsdictionary.com/dictionary/en..> is with the lowest value of k. Hence the URL having k value 3.404 ranked first, and the with the k value 0.246 the URL ranked as last.

GRAPH SHOWING COMPARATIVE STUDY OF WEB SEARCH ENGINES RANKING

We have found that when we plot a graph between time to dataset size, it shows the time taken by the algorithm to the size of data it is working on. We plotted the graphs initially when previous database values are also been taken into consideration.

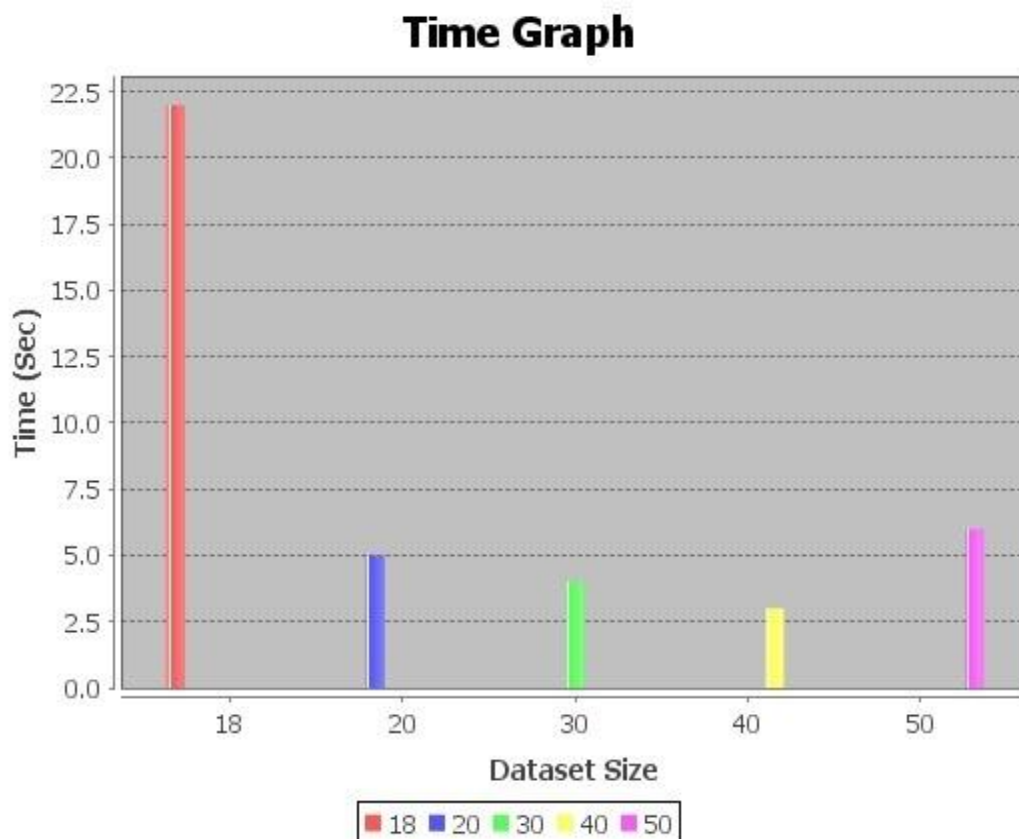


Fig. 12 DATASET SIZE VS TIME (SEC)

Figure given below shows that the whole procedure actively depends on the size of data, and it takes considerable time when the data size is very large. But shows an efficient time complexity when a relevant searching done.

PRECISION COMPARISION OF EXISTING VS PROPOSED

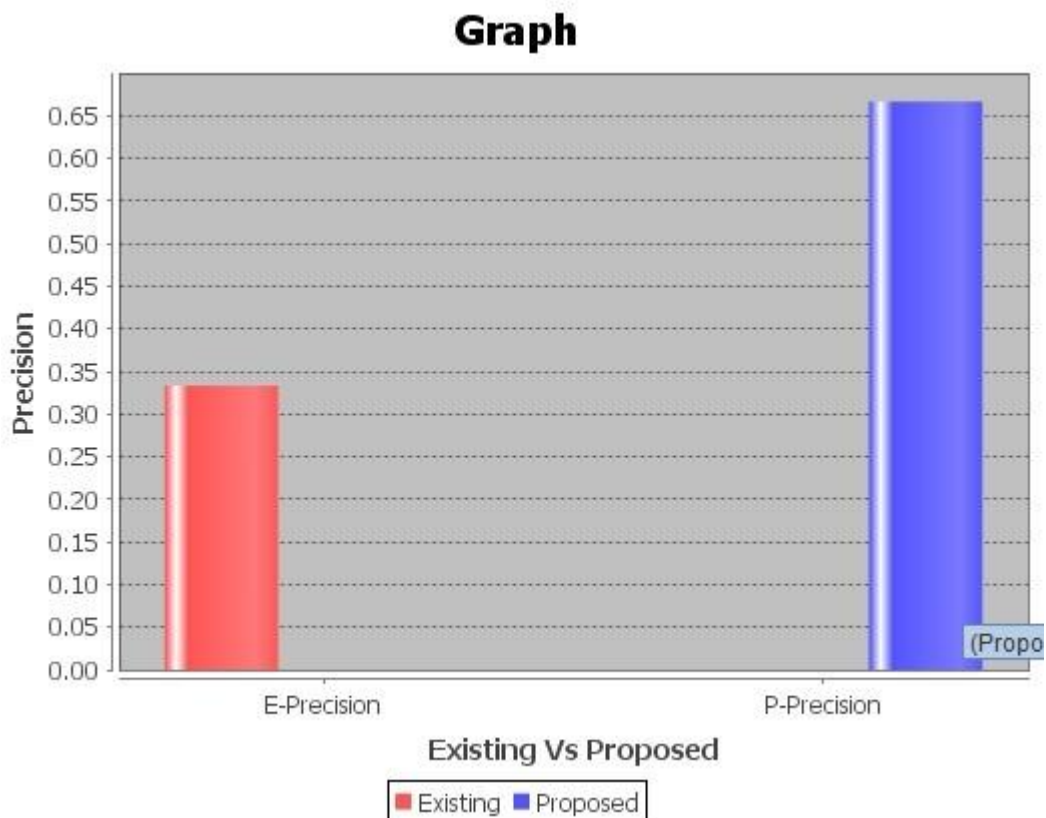


Fig. 13 PRECISION COMPARISION OF EXISTING VS PROPOSED

Above figure is a comparison graph of the performance of search engines when simply the existed or content based ranking of web pages done, to the proposed approach where we have focused on clustering techniques to classify the web documents based on their relevancy.

CHAPTER 5: CONCLUSION AND FUTURE DIRECTIONS

In this era of internet, the search engines are very efficient tool for information retrieval. Therefore, the search engine has to perform according to the mark and requirement. As the data reside online we have a very huge of data at every time on online data resources. This makes the availability easy but on the other hand the fetched data is so inconsistent and noisy that there has to be an efficient method to keep the data in an organized manner. Putting them in an organized way so whenever a user requests any information in the form of query he/she will get what they want in the form of web documents. According to the given approach and work we can say that the this work is producing high accuracy in terms of precision when we rank the web documents compared to other search engines result.

In future we can also get better results if we perform tuning operations on the search results. Many machine learning and data mining techniques making tasks easy. Here we are only concentrating on text data and ranking methods based on it to get the relevant and noise free search results. However these days the data is available in various forms like images, audio and videos.

CHAPTER 6: REFERENCES

- [1] Bing Liu, Kevin Chen- Chuan Chang ,” Editorial: Special issue on Web Content Mining” , *SIGKDD Explorations*, Volume 6, Issue 2.
- [2] Bin W, LiuZhijing, Web Mining research, *5th International Conference on computational Intelligence and Multimedia Applications*, 2003
- [3] Brin, S., and Page, L., 1998. “The anatomy of a large-scale hyper textual Web search engine”, *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp: 107-117.
- [4] Chakrabarti, S. “Mining the Web: Discovering Knowledge from Hypertext Data”, *Morgan-Kauman Publishers*,2002.
- [5] Chakrabarti, S., Berg, M., and Dom, B. Focused Crawling: A New Approach to Topic specific Web Resource Discovery. *Computer Networks, Amsterdam, Netherlands*, 1999.
- [6] Cheng Wang, Ying Liu, Liheng Jian, Peng Zhang, A Utility based Web Content Sensitivity Mining Approach, *International Conference on Web Intelligent and Intelligent Agent Technology (WIAT), IEEE/WIC/ACM 2008*.
- [7] R. Cooley, B. Mobasher, and J. Srivastava. “Web mining: Information and pattern discovery on the World Wide Web”, *In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’97)*, 1997.
- [8] R. Cooley, B. Mobasher and J. Shrivastava, “Web Mining: Information and Pattern Discovery on the World Wide Web”, *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, pp. (ICTAI), 1997.
- [9] R. Kosala, H. Blockeel, “Web Mining Research: A Survey”, *SIGKDD Explorations*, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.
- [10] I. Mele, “ Web Usage Mining for Enhancing Search – Result Delivery and Helping Users to Find Interesting Web Content,” *ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’13)*, pp. 765-769, 2013.
- [11] P. Sudhakar,G. Poonkuzhali, R. Kishor Kumar, “Content Based Ranking for Search Engines”, *Proc. International Multi Conference of Engineers and Computer Scientists (IMECS 12)*, 2012.
- [12] Georgies Lappas, An overview of web mining in societal benefit areas, *The 9th IEEE International Conference on E-Commerce Technology*, IEEE 2007.
- [13] U. Lee, Z. Liu, and J. Cho, “Automatic Identification of User Goals in Web Search,” *Proc. 14th Int’l Conf. World Wide Web (WWW ’05)*, pp. 391-400, 2005.
- [14]O. Zamir and O. Etzioni, “ Web Document Clustering: A Feasibility demonstration,” *ACM (SIGIR, 99)* , pp. 46-54.

- [15] P. Chahal, M. Singh and S. Kumar “*Ranking of Web Documents using Semantic Similarity*” Information Systems and Computer Networks (ISCON), 2013 International Conference on 2013 IEEE, DOI 10.1109/ICISCON.2013.6524191 Page(s): 145 – 150
- [16] G. Kumar, N. Duhan and A. K. Sharma “*Page Ranking Based on Number of Visits of Links of Web Page*” Computer and Communication Technology (ICCCT), 2011 2nd International Conference on 2011 IEEE, DOI 10.1109/ICCCT.2011.6075206 Page(s): 11 – 14
- [17] P. Rani and Er. S. Singh, “*An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters*” INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY Vol. 9, No 1 J u l y 1 5, 2 0 1 3
- [18] A. Jain, R. Sharma, G. Dixit and V. Tomar, “*Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages*” Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 2013 IEEE, DOI 10.1109/CSNT.2013.137 Page(s): 640 – 645
- [19] N. Batra, A. Kumar, Dr. D. Singh and Dr. R.N. Rajotia “*Content Based Hidden web Ranking Algorithm(CHWRA)*” Advance Computing Conference (IACC), 2014 IEEE International, 2014 IEEE, DOI 10.1109/IAdCC.2014.6779390 Page(s): 586 – 589
- [20] H. Dubey, Prof. B. N. Roy “*An Improved Page Rank Algorithm based on Optimized Normalization Technique*” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5), 2011, Pages(s): 2183-2188
- [21]N. V. Pardakhe, Prof. R. R. Keole “*Analysis of Various Web Page Ranking Algorithms in Web Structure Mining*” International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013
- [22] N. Höchstätter and D. Lewandowski, What users see – Structures in search engine results pages, Information Sciences 179 (2009), 1796–1812.
- [23] M. Steinbach, G. Karypis and M. Kumar, A Comparison of Document Clustering Techniques, KDD Workshop on Text Mining (2000).
- [24]M. Alam and K. Sadaf, A Review on Clustering of Web Search Result, Advances in Intelligent Systems and Computing Volume 177 (2013), 153-159.
- [25] O.Zamir and O. Etzioni, Web document clustering: A feasibility demonstration. In Research and Development in Information Retrieval (1998) ,46-54.
- [26] G. Mecca, S. Raunich and A. Pappalardo, A New Algorithm for Clustering Search Result, Journal of Data & Knowledge Engineering, Volume 62, Issue 3 (2007).
- [27] Y. Wang and M. Kitsuregawa, Link Based Clustering of Web Search Results, In Proceedings of The Second International Conference on Web-Age Information Management (WAIM2001), Xi'An, P.R.China, Springer-Verlag LNCS (2001).

- [28] Mercy paul Selvan,ChandraSekar ,A.Priya Darshan Survey on web page ranking algorithms in ijca (International Journal of Computer Applications) proceedings of 2012
- [29] Shital C Patil,RR keole Content and usage based ranking for enhancing search engine delivery 2014 in volume 3 issue 7International journal of Science and Research (IJSR)
- [30] “Metasearch engines and Information retrieval :Computational Complexity of ranking multiple search results” at fifth International Conference on Information Technology: New Generation 2008.1