ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

Strong Information Scent based Search Engine Design

Piyush Kumar[1], O.P Verma[2]

P.G Student, Department of Information Technology, Delhi Technological University, New Delhi, India[1]

Head, Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India[2]

ABSTRACT: "Wealth of information creates the poverty of attention" ~Herb Simon.WorldWideWeb is the grid of valuable information which is extending with the enormous speed. World Wide Web consists of billions of web pages and their urls and thousands of web pages keeps adding every day. Search engines provide the way to access these web pages and their links based on the query fired by the users and in return, search engine provides users with the several links of the web pages which may consist of the user's information need. But, due to the ample of links provided by the search engine in response to the fired query, user may get confused which link to follow, which will contain his/her information need. This paper proposes a novel optimization design for search engine which considers the web log mining of the search engine query log, also called as click through data and the information foraging theory concept i.e. information scent to provide the user with the more optimized links of the web pages in order to reduce their information snacking time.

KEYWORDS: Web Log Mining, Information Foraging, Information Scent, Clustering, Page Rank, Cosine Similarity

I. INTRODUCTION

We are living in the era of information age. Every day, multi-set of web pages ,terabytes or petabytes of data drain into the computer networks, world wide web and various storage devices from science and engineering, medicine, business societies and almost, from every aspect of daily life

This explosion of the available data volumes is the repercussion of the computerization of our society and rapid development of the powerful data collection and storage tools. The catalogue of the data sources that generate the ample amount of data is endless. Hence, there is an indispensible need of the powerful and versatile tools which can uncover the valuable information from the tremendous amount of data and to transform such data into organized knowledge. This necessity has led to the need of web search engines. A web search engine is a specialized computer server that hunts for the information from the web. It receives millions of queries every day. Each query can be viewed as the transaction where the user describes his or her information need. Web search engines are actually very large web mining applications. It uses various web mining techniques like web content mining, web log mining, web structure mining and web crawling[1],[2],[3].But, web search engine poses grand challenges to web mining. It has to handle a huge and ever-growing data and returning the result from such a huge data source is very tedious and over that returning the optimized result is even more fugacious.

There are many searching techniques of the search engine such as Page Rank algorithm[4],HITS (Hypertext Induced Topic Search)[4] and Weighted Page Rank algorithm[5] which has improved the searching from, such a vast amount of web data, to a very large extent. These algorithms find the rank of each page which defines the quality of a page and use these ranks to return the web results to a user in response of their queries. The results are returned in the decreasing order of their ranks. But still, the problem is that the user will get large number of web pages in return to their queries. The users information may be satisfied by the top rank search engine results or may not be. Eventually, the user has the large number of returned web pages but user may not be able to decide which link to follow. He/she may follow different-different links of returned web pages and probably won't get the required information. The useralso may get confused which link to navigate when addressed by the search engine web results. This only increases the user's navigation time to fetch the information.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

This paper describes the novel approach to return the subset of the web search result of the search engine in which the probability of satisfying the user's information need to a great extent. This lead to save the user's navigation and information snacking time.

The paper is organized as follows: Section II describes the related work and concepts which are used in proposed work. Section III describes the proposed architecture of search engine, proposed similarity metric and the clustering algorithm. Section IV discusses the practical evaluation of the work. Section V concludes the paper.

II. RELATED WORK

(A) Web Log Mining

User's activities are stored in the log file by almost all the search engines. This log file contains the information about the users such as user's id, query fired by user, clicked urls, rank r of the clicked urls u for query q, time at which query is submitted to the search engine. This log file is also called as click through data.

Web log mining is one of the techniques of web mining which aims at mining the log file in order to gain insight information about the user and how it is using search engine and what are his or her interests which, in turn, can be used to recommend suggestions about the queries and web pages to the user. Web log mining can also be used to find the similarities between queries and between the clicked urls whichgives the necessary information to provide recommendations to the users.

(B) Information foraging

Information foraging [6],[7] is the theory that applies the concept of optimal foraging theory to the idea about how human users search for information.The theory is based on the assumption that, when humans searching for information, humans use "built-in" foraging mechanisms in order to forage as much as relevant information as possible like animals find food and grab as much as possible. The theory assumes that user, when possible, will modify their strategies or the structure of the environment to maximize their rate of gaining valuable information [8].The information foraging can be summarized, as shown in fig.1, as the psychological behavior of humans who uses the optimal foraging technique for fetching the information from the web.

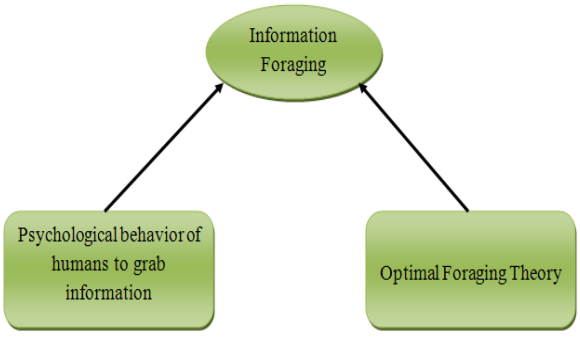

Fig. 1: Information Foraging

1.Details of Theory of Information Foraging

- Organisms that consume information are called **informavores**. Informavores constantly makes decision on what kind of information to look for, whether to stay at the current site to try to find additional information or whether they should move to the other site, and which path (link) to follow to the next information site.

- In **optimal foraging theory**, our emphasis is on maximizing the energy per unit time.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

$$Optimal\ Foraging\ = maximize[\frac{ENERGY}{TIME}] \quad (1)$$

- In the same way, in **information foraging theory**, our emphasis is on maximizing the useful information per unit time

$$Information\ Foraging\ = maximize[\frac{USEFULINFORMATION}{TIME}] \quad (2)$$

- **Information scent[8],[9]** is a part of information foraging theory. It refers to the extent that users can predict what they will find pursuing a certain path through a website or other source of information.

(2) Information Scent

It is the most pivotal concept in information foraging. Like animals, rely on the scents which indicate the possibility of finding the prey and guide them to the promising patches, in the same way, web surfers rely on the various cues on the links in the web search area. Their surfing patterns are guided by these cues and their information needs.

It is the measure which tells whether the user who is surfing over the web finds the information that he/she looking forwhile clicking various links. It is of two types which are as follows:

- **Strong information scent:** It refers to the measure which tells that, when user presented with the list of options which option they will select that gives the clearest indication of the information they are looking for, and whether they find the required information after clicking this option.
- **Weak information scent:** It refers to the measure which evaluates the links which does not provide the clear indication of the user information need.Hence, confuses the user which link to follow.

Consider Table 1 which presents some set of navigation options to users while he/she browsing online shopping site. Each option is giving some information to the user i.e. what the option will contain. For example, 'Audio and TV' will contain electronics items, 'Books' will contain book library, 'Computing' will contain all the computer accessories and electronics items.

Table 1: Navigation options in online shopping site for evaluating information scent

---------Select options----
Audio and TV
Books
Computing

If user wants to look for information about i-pods which option he/she will choose here. Here, all the navigation options are clearly understandable, but which option user will select to gain information on i-pods? Since i-pods can be listed in either'Audio and TV' or 'Computing' section. In this case, user is more likely to select the wrong option which makes the user get frustrated on not finding information. The user backtrack the links or eventually leave the site. This is referred to as weak information scent.

III. PROPOSED SEARCH ENGINE DESIGN

The optimization technique used in the proposed search engine is based on mining the clicked through data in order to find the similar queries and common clicked urls and how much time is spent on the common clicked urls which is used to suggest the strong information scent based web pages to the users.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

The architecture of the proposed search engine as shown in fig 2 comprises of following components:

1. *Search Engine Interface*: User writes the queries at the search engine interface.
2. *Query Processor:* It processes the request of the query. It finds the Euclidean distance (D) between the query and the centroid of each cluster and fetches the urlsfrom the clusters, if Euclidean distance (D) ≥ Threshold. The fetched urls are returned according to their ranks set by rank analyzer.
3. *Query Log:* Query log is the log file which comprises of the database of all the queries fired by all users, user ID, time of query, clicked urls. The query log provides necessary details about the users which is used for mining and helps discovering their surfing pattern.

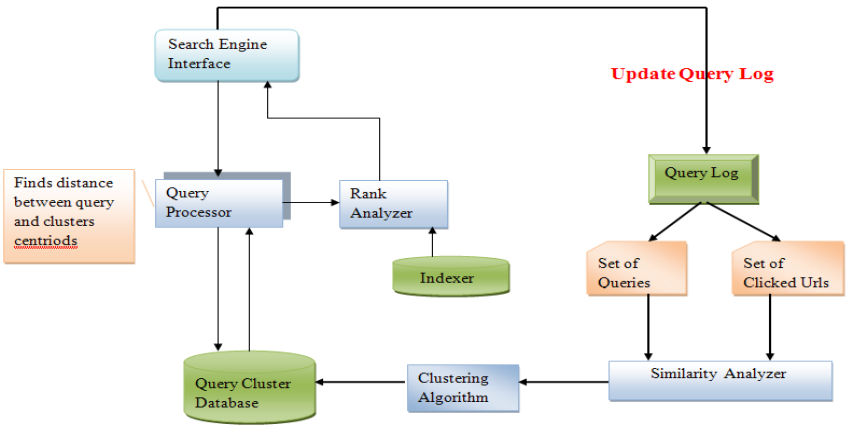

Fig 2: Proposed Search Engine Architecture

4. *Similarity Analyzer:* similarity analyzer will find the similarities among the queries of different users and similarities among the clicked urls after query is processed. In similarity analyzer, time spent on the clicked urls is also considered. It will be explained in later section in detail.
5. *Clustering Algorithm:* It will perform the clustering on the basis of found similarity. It will form the cluster of similar queries with their clicked urls. It will be explained in later section in detail.
6. *Query Cluster Database:*It will make the database of clusters of the queries with their urls and these clusters are the sources which return the strong information scent web pages after processing of user's query by search engine.
7. *Rank Analyzer:*Rank Analyzer will rank the urls fetched by the query processor from the query cluster database. It will work on the Page Rank algorithm.

The similarity analyzer and clustering algorithm are the crux of the proposed search engine, which are as follows:

(A) Similarity Analyzer

Similarity analyzer works on three similarity criteria:

- Similarity between queries.
- Similarity between clicked urls.
- Time spent on a clicked web page.

(1) Similarity between users queries

Similarity between the users queries in the log file is calculated by finding the common content words between the queries. This similarity is assisted by using the cosine similarity. It is given by

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

$$Sim_{query}(q_1, q_2) = \frac{q1.\ q2}{\|q1\|\,|q2|} \quad (3)$$

Where q_1 and q_2 are the two queries comprises of say 'z' number of words each. Here, $\|q_1\|$ is the Euclidean norm of query $q_1=(q_{1a},q_{1b},q_{1c},\ldots\ldots, q_{1z})$ where q_{1a} represents one word of a query q_1, q_{1b}represents following word of a query $q_1$and so on. Euclidean norm of query $q_1$defined as $\sqrt{{q_{1a}}^2+{q_{1b}}^2+{q_{1c}}^2+\ldots\ldots{q_{1z}}^2}$ .Similarly, $\|q_2\|$ is the Euclidean norm of query $q_2=(q_{2a},q_{2b},q_{2c},\ldots\ldots, q_{2z})$where q_{2a} represents one word of a query q_2, q_{2b}represents following word of a query q_2 and so on. Euclidean norm of query q_2 defined as $\sqrt{{q_{2a}}^2+{q_{2b}}^2+{q_{2c}}^2+\ldots\ldots{q_{2z}}^2}$.

(2) Similarity between Clicked Urls

Two queries are similar, if it lead to the selection of the same or similar web pages. It is given by [10],[11] i.e.

$$Sim_{clicked}(q_1, q_2) = \frac{\text{commom urls clicked in two queries(q1, q2)}}{\text{maximum(clicked urls in queries(q1, q2))}} \quad (4)$$

(3) Total Time spent on Common Clicked Web page between Two Queries

In the proposed methodology, the time spent on a web page is also considered to give more weight to the similarity measure. It is assumed that if a user is spending time which is greater than some threshold time, then, that page reflects strong information scent i.e. the quality of a web page is good where user's information need is satisfied.

(B) Proposed Similarity Metric

Proposed Similarity metric is given by:

$$S_m = Sim_{query}(q_1, q_2) + Sim_{clicked}(q_1, q_2) + \theta_o \quad (5)$$

Here,
$Sim_{query}(q_1,q_2)$ is Similarity between users queries
$Sim_{clicked}(q_1,q_2)$ is Similarity between Clicked Urls.
θ_o is added to the metric if Time spent on common clicked urls is greater than some threshold. The value of θ_o can be taken as any value to give more weight to S_m. Here, the value is considered 1, which is concluding good results.

(C) Clustering Algorithm

Clustering algorithm is used to cluster the query along with their corresponding clicked urls and time spent on these urls on the basis of the proposed similarity metric score. Following is the proposed clustering algorithm used:

Step1:

Input: set of n queries with their corresponding clicked urls and time spent on the clicked urls.
Minimum similarity threshold (£)
Output: set of clusters C= {C_1, C2.....C_p}, each cluster is stored in separate array.

Clustering_Query (n, £){
For (each query q_1 € n queries){
Set cluster_Id (q_1)= C_p
C_p= {q_1}
For(each query q_2 €{n-q_1} queries){
Find S_m

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

If($S_m \geq$ £) then{
$C_p = C_p U \{q_2\}$
}
Else
{ Continue }
}
p = p+1
}
Return C_p}

Step2:

Now, the mean value of each cluster, generated in step 1, is calculated which act as the mean centroid of each cluster The algorithm for finding mean centroid is as follows :

Input: set of clusters, C = {C_1,C_2.....C_p},where each cluster comprises of query content words, their common clicked urls and the time spent on each clicked url.
Output: Mean centroid of each cluster i.e. mean S_m, distinct keywords

Mean_Centroid_Calculate(C){
For (each cluster C_p€ C){
Find mean centroid i.e. mean $S_m$value of each cluster and distinct content words of queries in each cluster
Return mean centroid
}
}
Mean centroid of each cluster is calculated because when a user fire the query in the search box, the similarity(S_{query}) between the fired query and the distinct content words in each cluster is calculated and then, Euclidean distance(D) between the mean S_m(centroid) and Squery is calculated. If D $\geq$ threshold, the urls from the corresponding cluster (clicked urls in each cluster are strong information scent based urls) are fetched and shown to the users according to their ranks. This whole clustering process discussed here, is highly iterative process.

IV. EXPERIMENTAL RESULTS

The proposed work is practically evaluated on the query log as shown in table 2(table 2 is a sample segment of query log). The proposed similarity metric(S_m) and the proposed clustering algorithm is applied on table 2 which is as follows

Table 2: Query Log File

RowId	Sno	Query	UserId	Clicked Url	Time Spent(minutes)
1	1	web mining	123	http://www.xyz.com	1
2	1	web mining	456	http://www.abc.com	5
3	2	data mining over web	789	http://www.abc.com	10
4	2	data mining over web	189	http://www.xyz.com	5
5	3	web data mining	198	http://www.pqr.com	3
6	3	web data mining	123	http://www.rst.com	8
7	3	web data mining	201	http://www.lmn.com	10
8	4	mining example	198	http://www.xyz.com	5
9	4	mining example	111	http://www.xyz.com	6

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

(A) Proposed Similarity Metric Evaluation

Consider Sno 1 query and Sno 2 query, say:
q1: web mining
q2: data mining over web

Step 1: finding $Sim_{query}(q_1,q_2)$ as given by eq. (3).

*Step 2:*finding $Sim_{clicked}(q_1,q_2)$ as given by eq. (4)

Step 3: Finding total time spent on the clicked urls common to q1 and q2. If total time is greater than some threshold value add +1 for each set of common clicked urls for queries q1 and q2.

Step 4: Summing the result of step1, step 2 and step 3 to obtain proposed similarity metric.

Performing these steps for each pairs of queries i.e. Sno 1-3,Sno 1-4, Sno 2-3, Sno 2-4, Sno 3-4. The result for similarity metric evaluation is given by table 3.

Table 3: Proposed Similarity Metric Result Table

q1	**q2**	**S_w(Proposed Similarity Metric)**
(1) Web mining	(2)Data mining over web (3)Web data mining (4)Mining example	3.707 0.816 2
(2) data mining over web	(3) web data mining (4) mining example	0.866 1.853
(3) web data mining	(4) mining example	0.408

(B) Clustering

Step 1: The clustering algorithm defined in section III is applied to the table 3 following clusters will be formed and considering S_m threshold is 1:
C_1: {(1),(2),(4)}
C_2: {(2),(4)}
C_3: {(3)}

*Step 2:*Mean centroid of each cluster is calculated by using the algorithm defined in Section III.

(C) Comparing Clustering with different threshold values

Different clusters will be formed with different threshold values. The result with varying threshold values are listed in fig 3, 4 & 5.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

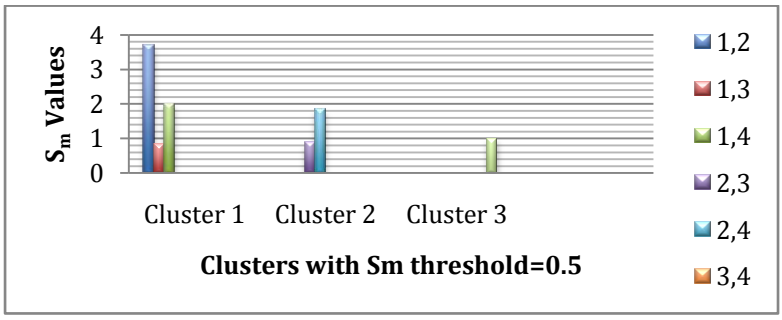

Fig 3: Clusters formation with S_mThreshold = 0.5

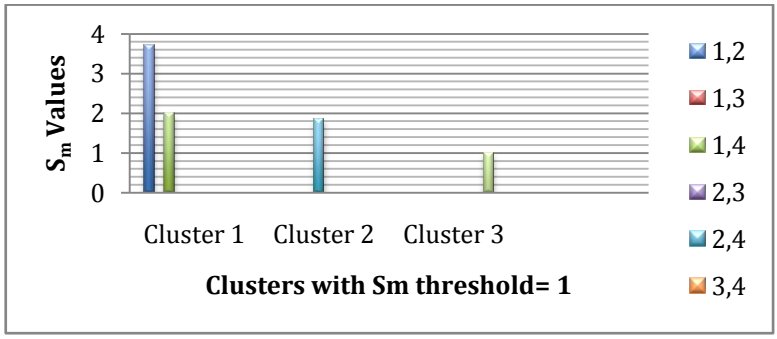

Fig 4: Clusters formation with S_mThreshold = 1

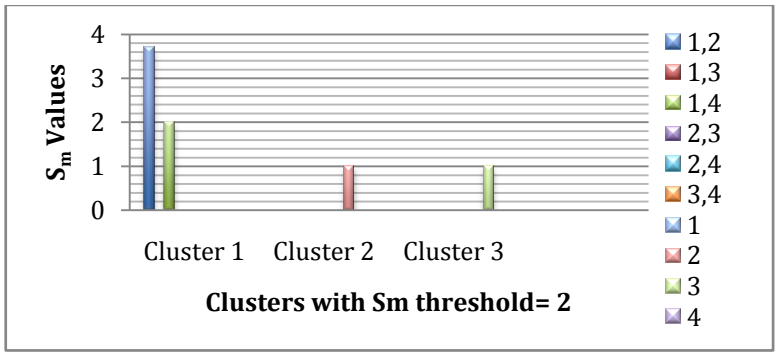

Fig 5: Clusters formation with S_mThreshold = 2

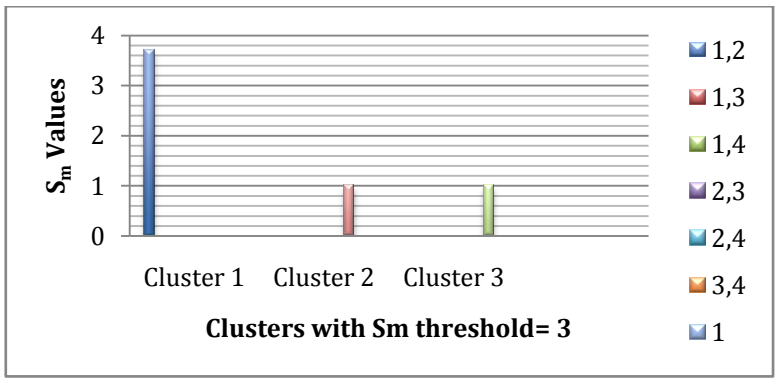

Fig 6: Clusters formation with S_mThreshold = 3

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

V. CONCLUSION

Search engines, indeed, the best application of web mining which returns the links in return of the query. Previous research works on optimizing the result of search engine are based on the page rank algorithms, weighted page rank algorithm, Hypertext Induced Topic Search (HITS) which has improved the result of search engine to a large extent. But still, users gets huge list of links in response to their queries which may confuses the user which one to navigate and increases the information seeking time of the surfer. The current work embeds the web mining techniques with the theory of information foraging and focuses on how to optimize the web search result in order to reduce the surfer's navigation time and returning the user with the web pages which satisfies his/her information need. The proposed work makes use of web log mining, clustering, cosine similarity and concept of information scent to accomplish the task of providing the user with the strong information scent based web pages to the surfer. The proposed work is assisted by the experimental results which conclude the result with the formation of query clusters along with their common clicked urls and time spent on these urls.

REFERENCES

[1] JaideepSrivastava, PrasannaDesikan and VipinKumar,"Web mining- Accomplishments and Future Directions".
[2] RenataIvancsy and IstvanVjak, "Frequent Pattern Mining in Web Log Data", ActaPolytechnicaHungaria, Vol. 3, No. 1, 2006.
[3] Raymond Kosala and HendrikBlockeel,"Web Mining Research: A Survey ", ACM SIGKDD, Volume 2 Issue 1, july 2000.
[4] Joni Pajarinen, "Page Rank/HITS algorithm ", March 19, 2008.
[5] Shesh Narayan Mishra, AlkaJaiswal, AshaAmbaikar,"An Effective Algorithm for Web Minining based on Topic Sensitive Link Analysis", IJARCSSE, Volume 2 Issue 4, april 2012.
[6] JefreeyHeer, Ed. H. chi,"Identification of Web User Traffic Composition using Multi-Modal Clustering and Information Scent", Xerox Pero Alto Research Center, CA 94304, 2000.
[7] Peter Perolli and Stuard K. Card,"Information Foraging ", Psychological Review, UIR Technical Report, January 1999.
[8] Peter Perolli, "Information Foraging and Information Scent: Theory, Models and Applications", Xerox Pero Alto Research Center.
[9] Ed. H Chi, Peter Perolli, Kim Chen and James Pitkow,"Using Information Scent to Model User Information Needs and Actions on the Web", Xerox Pero Alto Research Center, CA 94304, ACM, 2001.
[10] A.K Sharma, NehaAggarwal, NeelamDuhan and RanjanGupta,"Web Search Result Optimization by Mining the Search Engine Query Logs", IEEE,2010.
[11] R. Umagandhi, Dr. A.V.SenthilKumar,"Time Independent Query Recommendations from Search Engine Query Logs",IEEE.