International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

A Survey On Page Ranking And Search Engine For Query Log

Ankita Tiwari[1], Sushil chaturvedi[2]
[1, 2,] *PROF. Dept. of . (CSE/IT), S.R.C.E.M College, Gwalior (India)*

Abstract: Ranking methods have come a long way in past years with the beginning of the space vector models, various feature based approaches been proposed. Search engines are allowing the individuals to specify the queries simply as the keywords lists, following approach of the retrieval of the traditional information type systems. SEO is way of improving visibility of website or a webpage in the search engine by natural search or unpaid searched results. In optimization of search engine updating or modification of all variables to get a better location in search engine take place. Recommending the users with most useful keywords search not only to enhances the search engine's hit rate, but helps individuals to find the desired type of information more instantly. Behind such types of query the recommendation also interfaces, the important role is played by query log study. Clustering is a learning which is called an unsupervised not supervised learning like Classification. In the clustering method, dataset objects are then grouped into the clusters, in a way that the groups are different from one other ,the objects which are in same group or the cluster are identical to the each other.
Keywords: Web page ranking, Search engines, Search engine optimization (SEO), Query log, Clustering,etc.

I. INTRODUCTION

Web page ranking has now been based on hand designed by ranking function like BM25. However the ranking now being considered supervised learning problem and the various algo which is machine learning being applied to it.

Traditional technique in the learning to the rank optimizes uniform ranking measures such as number of mis-ordered pairs. However, it is often case that the users may be more interested in the most useful items (first page) and ignore other items. Thus it's appropriate more for a ranker to spend effort and get topmost items right. Performance measures have developed in information retrieval of community which pays more attention the most ranking. Examples of this type of measures are Normalized Discounted Cumulative Gain which is also called as (NDCG), Mean Reciprocal Rank this is also called as (MRR). They are needed to address problem of the evaluating the rankers, the search engines or recommender systems.

Ranking methods have come a long way in past years. Beginning with the space vector models, various feature based approaches been proposed. Popular set features include BM25 or its variants. Following the aim to show when adding such type of methods with the machine learning, the ranker performance can increased significantly.

The issues of ranking collections of objects d, such as webpages, products, movies, such that the popular objects are retrieved at beginning of the list. Typically we will do so given a query q. For ex, we need to sort collection of scientific papers those are related to query 'machine learning' are retrieved first.2 The basic belief we make is that such a ranking could be obtained by designing a scoring function g(d, q) which tells us how relevant document d is for query q.

The essentiality of able to tackle with each (document, query) pair independently arises from reasons of practicality. To find through a huge documents collection efficiently it is superior to able to score individually each document, in a particular when operations are performing in a distributed setting.

Users are often only interested in the most useful documents rather than the entire ranked list. For instance, for the web search, it is likely that users will only want to look at the first 10 retrieved results. Similarly, when retrieving documents, a user may only be willing to consider viewing the best k documents. Alternatively, satisfaction with the system may depending on how many documents he needs to sift through until he finds a relevant one.

which is common for all of those measures is that the order of the retrieved objects matters and that in many of the cases we will care primarily about the retrieval of useful documents rather than all of them.

Our approach to ranking is to centered around learning the scoring function g(d, q) with desirable properties. This is in disparity to many past strategies in information retrieval which rely on ingenious engineering to get good scoring functions. Obviously, engineering is still has its place in this type of framework after the all, we are needed to able to obtain the good features of (document, query) pair that could be used to define such a scoring function. Hence, here we take the advantage of statistics learning

and the machine learning to guide us in finding a function to that is optimized for this purpose. The design of good features remains the privilege of data retrieval expert [1].

II. SEARCH ENGINES

A vast factor for popularity of today's Web search engines called a friendly user interfaces (UI) they use to provide. , search engines are allowing the individuals to specify the queries simply as the keywords lists, following approach of the retrieval of the traditional information type systems. Keywords which refer to the topics which are broad, to the terminology in a technical form , or even proper nouns which may or may not guide process of the search the useful documents.

Despite this type of the simple form of interaction of the mechanism proved successful for Web searching, a keywords list of is not better descriptor of data needs of users. It's not easy always for the users formulate the effective queries for the search engines. One problem for this is ambiguity that is keeps on arises in many terms in language. Queries having terms which are ambiguous may get documents which are't what individuals are searching . whereas , users are typically to submit small queries in search engine, and the short queries which are likely be uncertain. From study of log of the search engine which is popular , summarized that all most all queries are to short (around two terms for per query) and imprecise.

Users searching for same knowledgeable data may phrase their queries in different way. Often, individuals try different type of popular queries until they satisfied with the results. Now to formulate queries in a effective way, individuals need to be familiar with some specific type of terminology in a domain which is knowledgeable. This not always the case: users have little knowledge about the information they are searching, and the worst, they couldn't be unquestionable about the needs that to be search for. As ex., a tourist accessing for the ads of summer rentals in the Chile may or may not called that the vast majority of such ads in the Web are for apartments in Vi˜na del Mar, a popular beach in the central most part of Chile. In contrast, the local users have the competence to submit the queries or problems with the term Vi˜na del Mar, when they are looking for a location to spend their vacations. The idea to use these type of expert queries to help non-expert users.

To overcome problems, some of the search engines has implemented methods that to suggest the queries alternative to users3. Their aim is to help the users to identify alternative related queries in their search process. Typically, list of queries which are suggested is calculate by the processing query log of search engine, which stores the history of previously the submitted queries and the URL's chosen into their answers. A major problem that occur in this context is to how model the useful information needs associated to a query. Few models proposed in the previous work represent that a query as set of URL's clicked by individuals for query. This technique have the limitations as when this comes to similar queries identification, because 2 queries which are related may output the different type of URL's in first places their solutions , thus the inducing the clicks into the different URL's.

A. Goals of Search Engine

1) Quality-Means effectiveness can defined to have the essential document set for a query. Process the text and are store text which is statistics to enhance the relevance be used.

2) Speed-Means efficiency may be defined as a process queries from users as fast as possible For it specialized data structure should be used.

B. How web Based Search Engine Works?

Web based search engine that works by saving the information of many web pages, which they retrieve itself These pages retrieved by web crawler also known spider which is following in every link on the site (Figure 1[3])

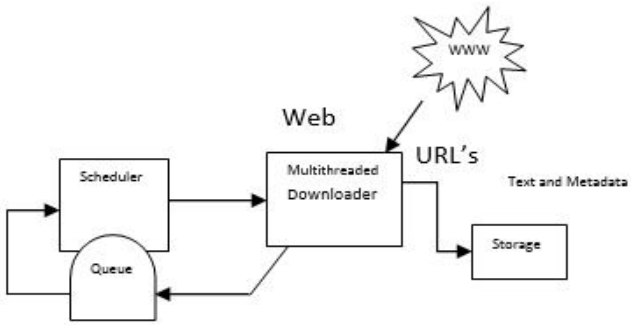

Fig. 1 search engine

Search engine used for information retrieval. Search engine match queries which are against index that are create by them. This index contains the word in the each document, pointers their type of location within the document. This is called inverted file.

C. Features of Web based Search Engine

Following are the basic features for evaluating web based search engine-

1) Web Indexes: When a web search request is generated. It is web index that are generated by the web robots or spiders. Combination of affects of the web indexes web search engine performance . 3 top most key points that are to design the web index are coverage, update frequency and the part of indexed web page.
2) Search Capability: Search Engine should provide Phrase searching; truncation Search capacity finds its performance efficiency, throughput.
3) Retrieval Issue: This issue proceed on three Key points- Precision, Recall and response time
4) Write Option: Write option or output option provides the deal with actual content of output.
5) User effort: User effort is about the documentation and the interface. Good prepared documentation and good interface play a different role in users' selection of web search engine. User wills only user the search engine when the interface is user friendly only.

D. Quality of Good Search Engine

1) Ability produce most useful result to any of the given search.
2) A true type of search engine is automated software program that moving around web collecting Web pages that to include in its catalog or the database.
3) It is searches when the user requests for the knowledgeable data from search engine having own the catalog or database of the collected Webpages, so you will be getting different results. Hits by using various other search engines.

E. Problems Facing by Current Search Engines

1) Crawlers are not able to analyze the content of keyword in web page before they download it.
2) User use to submits their request for the retrieval of useful data without mentioning content where he otherwise desire.
3) Crawler use to treats the user to search the request in isolation.
4) There is a requirement to make separate other files for each other type of web document.
5) Augmentation is then required in HTML document.

F. Types of the Search Engine

According to 3 different types search engine the functioning are [3].

1) Crawler Based Search Engine: They produce their listings are automatically . Spider builds them. Computer algorithm ranks all pages. These various other type of search engines are heavy and often get a lot of information. For complex search it allows to search within the results of the previous search and permit you to refine search results.
2) Human Power Directories: These are designed by human selection means they depend on professional to create listings. These never contain full text or webpage they link to.
3) Hybrid Search Engine: These are different from the traditional text oriented search engine like Google or directly based searched engine such as the Yahoo in which operates by the each program by comparing sets of the metadata.

G. Search Engine Optimization

SEO is way of improving visibility of website or a webpage in the search engine by natural search or unpaid searched results. different types of Optimization that target of the search like the image search, the local search, the video search, the academic search, the new search and the , specific type of industry search in a vertical form .It also define process of the affecting visibility of website or the webpage in search engine.

In optimization of search engine updating or modification of all variables to get a better location in search engine take place. We start with the SEO (Search Engine Optimization) and also the how it is used formulate the Strategy of internet marketing and also Technical aspects of the SEO. [3]

1) Using SEO as a marketing strategy it can be described as a method of how to getting our website to rank higher in search engine as Google; Yahoo, Means that if the user is liking to find a list of the optimized keywords chances are the visitors see your site on first few places may be good.

2) Parameters for evaluating the SEO of websites- Page Rank- Page rank of the each page depends on page rank of pages pointing to it.
3) To augment our rank of site page some key ideas are the inbound links, the outbound links, the Dangling links, the domain and the names of the File and the broken links [3].

H. Search Engine Optimization Technique

Basically three techniques for the optimization of search engine are there [3].

1) Directory Submission: It's the important technique in Search Engine Optimization to create in coming links to a website through indexed page and category. Different directory provides free service to website. Directory submission request information regarding URL, title, keywords.
2) Generation of Keyword: All optimization of the search engine required few words to enhance the information which based on the words like this. Keywords should of your system on the subject. This process proceeding by different online tools like word tracker, yahoo keyword tool selector , Google Ad words.
3) Link Exchange: To start up any website for any business we need reciprocal the link exchange with other websites. It is the procedure to take place link on other website and other website place links on our site.

I. Tools of Search Engines Optimization

SEO tools are operators that optimize the functionality of the search engine.[3] Basics tools are-

1) Keyword Tool- Include keyword research tools, analysis of keyword density tool, and competitor analysis tool. It is needed for website classification and regulate keywords deployment columns. Eg – Keyword tool selector , external keyword tool
2) Link Tool- These types of tools include the link popularity the spider simulator, by which the ranking of website increased.
3) Usability Tool- This tool test pages display effects in different resolution, different operating system, and different browser. These include HTML and CSS validation, Firefox extension, and Page speed test.
4) Keywords Strategy- When choosing keywords, it must be related with products, area, service.

High duality incoming Link- Submit the website to search engine directories, finds websites to exchange links. In it import link, outbound links, internal link are used.

J. Criteria for site optimization

For a new website to be optimized for the given keywords need to have some technical issues checked.[3]

1) Meta descriptions or Metadata Keywords
2) Keyword analysis
3) Title Tags
4) Page content
5) Headlines Tag
6) URL structure and domain
7) Images Tag
8) Page Load time
9) XML site Map
10) Meta data using schema
11) Site map
12) Robot.txt
13) 404 error
14) Duplicate contents

III. QUERY LOG

For search engine providing the related queries users help them to find desired content more quickly. For this reason, many search engines started displaying related to find keywords at the bottom of result page. Examples include Yahoo's ''Also try'' feature, Ask Jeeves's ''Narrow Your Search'', ''Expand Your Search'', and ''Related Names'' features, and Amazon's ''Related Searches'', etc. The motive is give search engine users a comprehensive recommendation when they search specific topics. Our statistics1 to show hit rate of related keywords search is quite a bit high or more than 10%. Recommending the users with most useful keywords search not only to enhances the search engine's hit rate, but helps individuals to find the desired type of information more instantly.

Behind such types of query the recommendation also interfaces, the important role is played by query log study. That is, we can also get the recommendations of query by the mining type query logs of search engine, that will contain huge data on the past queries.. A soft type of relation matrix is formed to store relation between the consecutive queries that occur within same session that is submitted by an individual inside one visit. Queries which are being submitted one after other immediately receive maximal relation of value, while relation value of this type is to dampened for the queries which are apart from other during same search session [4].

IV. QUERY LOG ANALYSIS

Query log analysis includes filtering methods and some crawling mechanisms that could be evaluated on the basis of part of speech tagging. There are 2 main approaches in retrieving the user‘s previous work related particular URL. This will include mining URLs for the evidence of pages visited by the user in the past session.

A. Filtering of methods help to remove words from dictionary and forming the documents. A standard type of filtering type method is stopped to word filtering. The stop word type of filtering idea is to erase the words that is to bear little or no information of content, like articles, the conjunctions, and prepositions, etc.

B. Part-of-speech tagging (POS) determines the part of speech tag, e.g. the noun, the verb, an adjective, etc. for each term. Text chunking here is to aims the grouping of the adjacent words in sentence. An eg of chunk is noun phrase ―current account deficit‖. Word Sense Disambiguation that used to try to resolve ambiguity in meaning of a words or a phrases. ex. is _bank‘ which have – among others – the senses _financial institution‘ or the _border of a river or lake‘. Thus, instead of terms the specific meanings could be stored in the vector space representation. This leads to larger dictionary but helps to considers semantic of term in the representation. Parsing is to produce a full parse tree of sentence. From the parse, we can find the relation of each word in the sentence to all the others, and typically also its function in the sentence [5].

V. CLUSTERING

Mainly the Clustering is method that is to includes grouping of the similar type of objects into 1 cluster and the same type of cluster which is to includes data objects set chosen to minimize the some of the measure of the dissimilarity. Clustering is a learning which is called a unsupervised not supervised learning like Classification. In the clustering method, dataset objects are then grouped into the clusters, in a way that the groups are different from one other ,the objects which are in same group or the cluster are identical to the each other[2]. Unlike Classification, which the predefined classes set that presented, but in the Clustering where there are not any kind of classes that are predefined set which resulting clusters not before known execution of the clustering algo. Here these clusters are being extracted from dataset by grouping objects in it. Types of Clustering Algorithms.

A. Hierarchical Clustering Algorithm
B. Algo which is K-means Clustering
C. Self organization maps (SOM)
D. Density Based Clustering Algo
E. EM clustering Algo [6].

VI. LITERATURE SURVEY

Gabriel Urrutia (2016) here, representation of A new equivalent rank constraints used to design low-order of controller with the prescribed degree of the stability. We will solve problem of optimization involving linear form of inequalities in matrix and rank constraints. we address issue of intriguing a reduced type of order feedback control. We incorporate rank constraints to restrict the order of the unknown controller through an optimization problem. The optimization resulting framework gives us the possibility of formulating additional rank constrained problems for control design. We apply an equivalent rank constraint representation to reformulate the problem into another one that is again equivalent in a optimum sense which are global by nature. The resulting (global optimum equivalent) problem solved by using the nonlinear programming approaches. We also formulate two additional extensions of the real reduced order control problem: 1) maximization of the stability degree, given a controller’s order, 2) given a certain stability degree, minimize the order of the controller. This shows the flexibility of rank constraint representation to solve different control design problems. Finally numerical ex are to adorn the achievements of proposed method [7].

Luca Brav (2016) here , we are considering the ranking of feature issues , where, there is given set of the training instances, task is to associate score with features to assess relevance. Feature ranking is necessary tool for the decision support type of systems, and used as an auxiliary type of step of the feature selection to minimize the high dimensionality of the real world data. We concentrate on problems of regression by assuming process underlying generated data approximated by a continuous type function (for instance, feed forward type of neural network). We formally state notion of the relevance of feature by recommending a minimum zero-norm inversion problem of a neural network, which is a non-smooth, forced optimization problem. We employ concave approx of zero-norm function, and we define smooth, and the issues of global optimization that to solved so as to assess the relevance of the features. We present new method named as feature ranking process which based solution of instances of the problem named global optimization depending on the available training data. on both the artificial data the Computational experiments and the real sets of data are performed, and point out that proposed ranking feature method is a valid alternative to existing methods in terms of effectiveness. The obtained outcome show that method is the most costly in the terms of CPU time, and this may be a limitation in the solution of large-dimensional problems [8].

YANG Min (2016) here, we propose optimization algorithm which is a lower rank to reconstruct it under the sampled dynamic MRI, which consisting of searching a matrix having the minimum rank subject to the linear equality constraints. To solve unconstrained non smooth convex type of optimization problem, we develop a fast alternating direction named method that uses the nuclear-norm to enforce low-rank constraint. Into its Lagrangian form, to solve through an accelerated proximal gradient (gradient step plus proximal step for nuclear norm) algorithm. Numerical experiments that on simulated set of data show encouraging results with low computational time, even at high acceleration rate [9].

Shipra Kataria (2016) here, Due to the tremendous growth of internet over past few years, a large repository of data covering almost every area has been formed over the web and as a result of which users of search engine are facing lot many problems in retrieving the most suitable information out of it which is known as information overkill problem. The main cause of this problem is non-optimization of web pages. This paper present a way for the investigation of the transaction logs that get through search engines to optimize rank of web pages and then resulting into the topic/subject relevant and user suitable documents at top of result pages of search engine. The algo begins with the query logs which is maintained by search engine to get into the exact information need of users. Then, an approach is to search the similarity among the queries which is based on two silent features i.e. query keywords and clicked URLs. Further, the query cluster that making tool is used forming the clusters of the queries which is of same kind based on the value of combined type of similarity measure which is lying b/w 0 and 1. After that, relevancy tool finder helps to works onto URLs that are associated with the each query in these clusters to search their relevancy with respect to the query by eliminating the effect of the black hat and several other search optimization techniques. A sorting algo is therefore applied on to each cluster which is to arrange the all URLs in the increasing order of relevancy and further sequential type of pattern mining algorithm is applied on them to search sequential pattern which are frequently accessed. The outcome of the procedure is improved then by ranking again the pages of web with help of calculation of weight according to newly discovered sequential patterns and the earlier rank associated with web pages [10].

Licheng Zhao (2015), this paper propose a work using a function which is smooth robust loss functions that to formulate robust problem of the lower rank of optimization problem in presence of the outliers. The aim of the problem is to regain a low rank of matrix of data from the entries which are noisy . Our main contributions are i) providing two smooth robust loss type of functions to handle respectively 2 types of outliers, i.e., universal type of outliers with unknown statistical distribution and the sparse spike-like outliers; ii) an algorithm which is efficient doing parallel minimization instead of alternating update. results which are of Numerical type show that proposed algo is to obtain a better result or solution at a faster rate of convergence of it than the state-of-art algorithms [11].

Yipeng Liu (2015) this , deals with issues that some signals of EEG having not a good sparse type of representation and a channel processing not efficient computationally in the compressed type of sensing of the EEG signals which is multi-channel. optimization model with the L0 norm and the Schatten-0 norm just to enforce cosparsity and the structures of low rank in reconstructed EEG signals which is a multi-channel. Both global consensus and the global consensus convex relaxation optimization with the direction of alternating method of the multipliers are needed to compute model optimization . The value and quality of EEG signal multi-channel reconstruction is the improved in way of both the computational type of efficient complexity and accuracy . The method which are proposed is much better candidate than the previous sparse type of signal recovery methods for the compressed sensing of the EEG signals. This method enables the successful compressed sensing of the EEG signals method even when the signals having not good sparse representation. Using the compressed type of sensing would reduce consumption of power of the wireless system of EEG system [12].

Andr´e Uschmajew (2015) here, they present rank-adaptive type of optimization strategy which is for finding low-rank solutions of matrix optimization problems involving a quadratic objective functions. The algo combines greedy outer type iteration which increases the rank and a smooth Riemannian algorithm that further optimizes the cost function on a fixed-rank manifold. This type of strategy is not a especially a novel, which we will show that it can be interpreted as concern gradient descent algo or as simple warm starting type of strategy of an algo known as projected gradient algo on the variety of the matrices of the bounded rank. In addition, to our numerical kind of experiments which show that the planning is efficient for recovering full rank but highly ill-conditioned matrices that have small rank of numerical type [13].

Xiaohang Chen (2014) here , a technique which is called Non orthogonal multiple access (NOMA) is promising radio type of access approach for the further cellular type of enhancements toward mobile communication which is of 5G systems. Single type user the multiple input multiple type of output (SU-MIMO) key technologies (LTE)/LTE- Advanced systems. Now proved that NOMA is combined with the SU-MIMO approaches can achieve further system performance improvement. Here, we focus on impact of the rank optimization on performance of NOMA with SU-MIMO in downlink. Firstly, a method named a geometry based of the rank which is method called adjustment method is studied. Secondly, an enhanced method for feedback rank adjustment is discussed. The results of simulation show that performance gain NOMA improves with the proposed two rank adjustment methods. Compared to the system known as orthogonal access system, large performance gains can be achieved for NOMA, which are about 23% for cell average throughput and 33% for cell-edge user throughput [14].

Pragya Kaushik et. al (2014) here , a method is to make the process of searching faster inside search engine. Cache support the search engine which refers to give a list of most expected type of queries which is possibly be entered by the individual after query submitting . These expected queries are associated queries with query given by the user. So here we had proposed a process to trace out associated queries and given an algorithm for this [15].

VII. CONCLUSION

In this paper, we perform the detailed survey on the various web page ranking techniques to show that how to rank the page. Basically search engines are used fetch the query related data. For the optimal web page ranking, search engine optimization has performed to increase the webpage visibility. Various amount of data are collected by the web page ranking then clustering performed to group them into useful information. To illustrate the useful concept of web page ranking, we provide the detailed survey.

REFERENCES

[1] Quoc V. Le, Alex Smola, Choon Hui Teo, “Optimization of Ranking Measures”, Journal of Machine Learning Research 1 (2999) 1-48 Submitted 4/00; Published 10/00
[2] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza, “Query Recommendation using Query Logs in Search Engines”, 49993cc7245969c6ed16bd0c233593c88cfe
[3] Rajesh Singh, S.K. Gupta, “Search Engine Optimization - Using DataMining Approach”, ISSN 2319 – 4847/Volume 2, Issue 9, September 2013.
[4] Zhiyong Zhang, Olfa Nasraoui, “Mining search engine query logs for social filtering-based query Recommendation”, Applied Soft Computing 8 (2008) 1326–1334.
[5] Chinmay R. Deshmukh, Prof. R .R. Shelke, “ URL Mining Using Agglomerative Clustering Algorithm”, ISSN: 2277 128X/ Volume 5, Issue 1, January 2015
[6] Supreet Kaur, Usvir Kaur, “ An Optimizing Technique for Weighted Page Rank with K-Means Clustering”, ISSN: 2277 128X/ Volume 3, Issue 7, July 2013
[7] Gabriel Urrutia, Ramon A. Delgado and Juan C. Aguero, “Low-order Control Design Using a Novel Rank-constrained Optimization Approach”, 978-1-922107-90-9 © 2016 IEEE
[8] Luca Bravi, Veronica Piccialli, and Marco Sciandrone, “An Optimization-Based Method for Feature Ranking in Nonlinear Regression Problems”, 2162-237X © 2016 IEEE
[9] YANG Min, “Low-Rank Optimization Algorithm for Accelerated Dynamic MRI”, 978-1-4673-9714-8/16/$31.00_c 2016 IEEE
[10] Shipra Kataria, Pooja Sapra, “A Novel Approach for Rank Optimization using search Engine Transaction Logs”, 978-9-3805-4421-2/16/$31.00_c 2016 IEEE
[11] Licheng Zhao, Prabhu Babu, and Daniel P. Palomar, “Robust Low-Rank Optimization for Large Scale Problems”,978-1-4673-8576-3/15/2015IEEE
[12] Yipeng Liu, Maarten De Vos, Sabine Van Hufel, “Compressed Sensing of Multi-Channel EEG Signals: The Simultaneous Cosparsity and LowRank Optimization”,
[13] Andr´e Uschmajew, Bart Vandereycken, “Greedy rank updates combined with Riemannian descent methods for low-rank optimization”, 978-1-4673-7353-1/15/$31.00 ©2015 IEEE.
[14] Xiaohang Chen, Anass Benjebbour, Yang Lan, Anxin Li, Huiling Jiang, “Impact of Rank Optimization on Downlink Non-Orthogonal Multiple Access (NOMA) with SU-MIMO”, 978-1-4799-5832-0/14/$31.00 ©2014 IEEE
[15] Pragya Kaushik, Sreesh Gaur, Mayank Singh, “Use of Query Logs for Providing Cache Support to the Search Engine”, 978-93-80544-12-0/14/$31.00_c 2014 IEEE.