ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Keyword Query Suggestion Using Document Proximity

Kajal P. Davange, Prof. P. N. Kalavadekar

Department of Computer Engineering SRES COE, Kopargaon, India

ABSTRACT: Query (keyword) recommendation in internet search helps user to find required information more quickly and accurately which improves the user search experience. Keyword suggestion have been used to recommend related queries. In keyword query suggestion using document proximity system directed weighted bipartitie keyword document graph is con- structed that captures the semantic relevance between keyword and distance between the user required document and the user location. Hospital dataset is used from this dataset system can find the nearest hospital. According to speciality of hospital system extract nearest three hospitals also display the distance and route of hospital from the user current location.

KEYWORDS: query suggestion, spatial databases,document proximity.

I. INTRODUCTION

Keyword suggestion (also known as query suggestion) has become one of the most essential features of business web hunt engines. Then afterward submitting a keyword query, the client may not be fulfilled with the outcome, so the keyword suggestion module of the search engine prescribes a z set of m keyword queries are generally to the users scan in the right manner [1].Viable keyword suggestion strategies are based on click majority of the data from inquiry logs and inquiry session data, or inquiry subject models new keyword suggestions can be determined according to their semantic importance to the original keyword inquiry. The previous techniques give location-aware keyword query suggestion (LKS), such that the recommended queries retrieve documents not only related to the user information needs but also located close to the client area [2].Many operation of a spatial database are advantageous from multiple points of view over suitable data. For object, in a geography information system, range search can be utilization to discover all restaurants in a certain area, same time nearest neighbor retrieval can find the restaurant closest to a provided address. The previous keyword suggestion systems do not consider the areas of the clients and the inquiry outcomes. Clients frequently bring challenges in expressing their web scan needs they might not think those keywords. After pro- viding a keyword query, the client may not be fulfilled those outcomes. This data gather due to the popularity of spatial keyword search Google transformed a daily average of 4.7 billion queries in 2011 a considerable portion of which have local function and target spatial web objects.

II. REVIEW OF LITERATURE

In paper[1] author describe about a weighted keyword- document graph, which captures both the semantic relevance between keyword queries and the spatial distance between the resulting documents and the user location.The graph is browsed in a random-walk-with-restart design, to choose the keyword inquiries with the most noteworthy scores as proposals. To make system adaptable, creator propose a partition-based approach that beats the standard calculation by up to an arrange of size. The fittingness of our framework and the execution of the calculations are assessed utilizing real data. The result appears that the system can offer useful suggestions and that Dad beats the pattern algorithm significantly. In the future, arrange to encourage think about the effectiveness of the LKS system by collecting more information and designing a benchmark. In expansion, subject to the availability of information, adjust and test LKS for the case where the locations of the inquiry backers are accessible in the inquiry log. Finally,it is accept that PA can also be used to speed up RWR on general charts with energetic edge weights; to explore this potential in the future.

In paper[2] author describe about to enhance search query log analysis by taking into account the semantic properties of

DOI: 10.15680/IJIRCCE.2018.0606040

query terms. Here to begin with depict a strategy for extricating a worldwide semantic representation of a look inquiry log and at that point appear how can utilize it to semantically extricate the client interests. The worldwide representation is composed of a scientific categorization that organizes inquiry terms based on generalization/specialization (is a) semantic relations and of a work to degree the semantic separate between terms. At that point characterize a inquiry terms clustering calculation that is connected to the log representation to extricate client interface. The assessment has been done on expansive real-life logs of a well known look motor.

In paper[3] author specifies about learning to rank arises in many data mining applications, ranging from web search engine, online advertising to recommendation system. In learning to rank, the performance of a ranking model is strongly affected by the number of labeled examples in the

training set; on the other hand, obtaining labeled examples for training data is very expensive and time-consuming. This presents a great need for the active learning approaches to select most informative examples for ranking learning; how- ever, in the literature there is still very limited work to address active learning for ranking. In this paper, explain about a general active learning framework, expected loss optimization (ELO), for ranking. The ELO framework is applicable to a wide range of ranking functions. Under this framework, author derive a novel algorithm, expected discounted cumulative gain (DCG) loss optimization (ELO-DCG), to select most informative examples. Then, to investigate both query and document level active learning for raking and propose a two- stage ELO-DCG algorithm which incorporate both query and document selection into active learning. Furthermore, shown that it is flexible for the algorithm to deal with the skewed grade distribution problem with the modification of the loss function. Extensive experiments on real-world web search data sets have demonstrated great potential and effectiveness of the framework and algorithms.

In paper[4] author describe a new indexing framework for location-aware top-k text retrieval and region-aware top-k text retrieval. The framework tightly integrates the inverted file for text retrieval and the R-tree for spatial proximity querying in a novel manner. Several hybrid indexing approaches are explored within the framework. The framework encompasses algorithms that utilize the proposed indexes for computing the top-k query, and it is capable of simultaneously taking into account text relevancy and spatial proximity to prune the search space during query processing. The conventional Inter- net is acquiring a geospatial dimension. Web documents are being geo-tagged and geo-referenced objects such as points of interest are being associated with descriptive text documents. The resulting fusion of geo-location and documents enables new kinds of queries that take into account both location proximity and text relevancy. This paper proposes a new index- ing framework for top-k spatial text retrieval. The framework leverages the inverted file for text retrieval and the R-tree for spatial proximity querying. Several indexing approaches are explored within this framework. The framework encompasses algorithms that utilize the indexes for computing location- aware as well as region-aware top-k text retrieval queries, thus taking into account both text relevancy and spatial proximity to prune the search space.

In paper[5] author specifies a formal treatment of the problem of query recommendation. In these framework author model the querying behavior of users by a probabilistic reformulation graph, or query-flow graph. A sequence of queries submitted by a user can be seen as a path on this graph. Assigning score values to queries allows us to define suitable utility functions and to consider the expected utility achieved by a reformula- tion path on the query-flow graph. Providingrecommendations can be seen as adding shortcuts in the query-flow graph that nudgethe reformulation paths of users, in such a way that users are more likely to follow paths with larger expected utility. Here in detail discussed about the most important questions that arise in the proposed framework. In particular, author provide examples of meaningful utility functions to optimize, then discuss how to estimate the effect of recommendations on the reformulation probabilities, address the complexity of the optimization problems that author consider, then suggest efficient algorithmic solutions, and validate models and algo- rithms with extensive experimentation. These techniques can be applied to other scenarios where user behavior can be modeled as a Markov process.

In paper[6] author discussed about a novel graph combination based rare query suggestion framework. The system divided into four major steps: 1. construct two query- URL bipartite graphs from query logs, where the click graph contains query-URL click information and the skip graph contains query-URL skip information, 2. perform random walk on each of the graphs, using the random walk with restart (RWR) technique, 3. build a correlation matrix for URLs from the category of URLs, 4. based on the URL correlation, iteratively optimize the model to estimate the best parameters of random walk and the combination rate of click and skip graphs. Finally,combine two query correlation matrices to form the optimal query correlation matrix, which is used for query suggestion. The model is inspired by

DOI: 10.15680/IJIRCCE.2018.0606040

pseudo-relevance feedback, it is valid to assume that search engines do not generate random top results. Users click the results based on their own perception of relevance so that the a URL may be clicked by one user but skipped by others. Ideally, all returned URLs should be considered relevant. However, since confident about top-ranked URLs, and only consider the skipped URLs above the last user click. Next,author discuss how the query- URL graphs are generated.
In paper[7] author specifies about experimentally study, query recommendations based on short random walks on the query-flow graph. The experiments show that these methods can match in precision, and often improve, recommendations based on query-click graphs, without using users' clicks. Ex- periments also show that it is important to consider transition- type labels on edges for having good quality recommendations. Finally, one feature is providing diverse sets of recommenda- tions: the experimentation that author conducted provides en- couraging results in this sense. The query-flow graph annotated with query reformulation types, can be used to generate query recommendations matching the ones obtained using query- click graphs. This means that the information contained in the annotated query-flow graph about consecutive queries in a session is as useful for this task as the user's clicks; given that both data sources are independent, recommendations produced by a composition of both methods are worth to be investigated as future work. When using the query-flow graph, author have found that it is important to discard edges between queries in different chains, even if they are frequent transitions. Also, allowing only certain reformulation types (e.g.: only specializations,or only specializations and parallel moves) is better than using the entire graph. Finally, doing a few iterations is better than doing one iteration (this is picking directly the node connected by the edge of highest weight), and more than 10 iterations in setting did not add precision to the results.

In paper[8] author describe about a novel query suggestion algorithm based on ranking queries with the hitting time on a large scale bipartite graph. Without involvement of twisted heuristics or heavy tuning of parameters, this method clearly captures the semantic consistency between the suggested query and the original query. Empirical experiments on a large scale query log of a commercial search engine and a scientific liter- ature collection show that hitting time is effective to generate semantically consistent query suggestions.The algorithm and its variations can successfully boost long tail queries, accom- modating personalized query suggestion, as well as finding related authors in research. In this paper, author discussed a unified approach to query suggestion, by computing the hitting time on a large scale bipartite graph of queries and click through. Despite its simplicity, this novel approach introduces quite a few benefits to query suggestion: 1) the suggestions generated with the proposed algorithm are semantically similar to the original query; 2) the suggestions generated do not have to occur with the original query; 3) this approach boosts the long tail queries as suggestions; and 4) this model provides a natural treatment for personalized query suggestion.
In paper[9] author specifies about data space is one of the emerging approaches to managing heterogeneous data sources. However, data spaces differ from conventional data integration approaches. Heterogeneous data sources with a pay-per-use in- tegration approach. When data spaces meet unstructured data, indexing is the only way to answer queries based on the best effort. As far as we know, we have not worked on the provision of relevant results for a query beyond those provided by the indexation without manual efforts. The approach proposed in this document stores. The relationships between the users queries. These relationships are used to improve the search results in the data space environment. The stored relationships are AND, OR and subsequent relationships (one keyword after the other). The proposed work is based on the way in which the search engines handle the same problem [9].

In paper[10] author introduces a technique for mining a collection of user transactions with an internet search engine to discover clusters of similar queries and similar URLs. The information is click through data each record consist of user users query to a search engine along with URLs which the user selected from among the candidates offered by the search engine. By viewing this dataset as bipartite graph,with the vertices to identify related queries and URLs. One noteworthy feature of the algorithm is that it is content ignorant. The discovered clusters to assist user in web search and measure the effectiveness of the discovered cluster in the Lycos search engine.

III. SYSTEM OVERVIEW

In keyword query suggestion using document proximity framework user fire a query which may be single word or phrase. Then from that input keywords are extracted using that keywords and documents i.e set of geodocuments with these two factors keyword document graph is constructed. It is directed weighted bipartite graph. Then next step is location aware edge weight adjustment. The edge weight adjustment is done based on loaction of the user query issuer and the node of the KD graph afterwords suggestion are recommended which depend on relevance of the keywords i.e initial user need and closeness of the document.

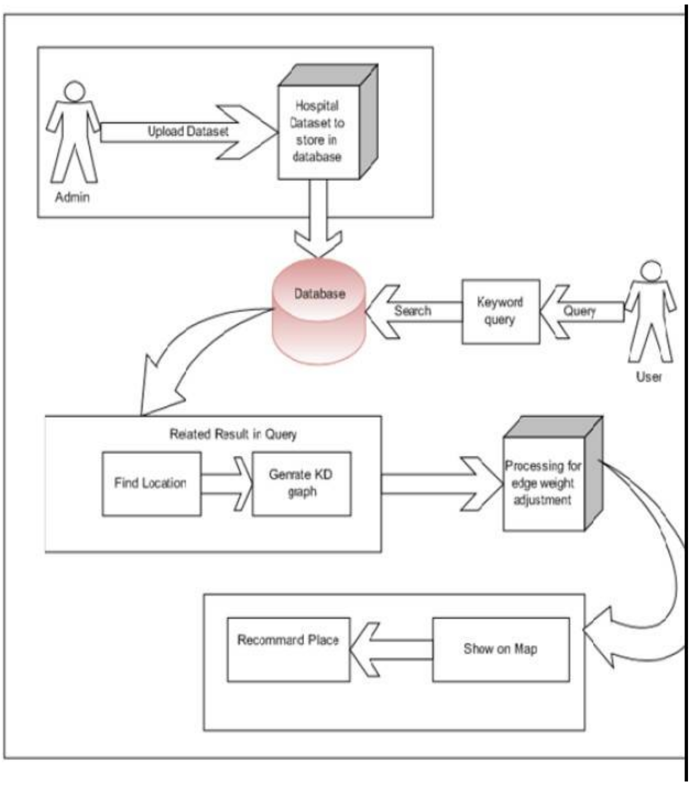


Fig. 1. System Architecture

Phase 1: Admin upload the hospital dataset. Then preprocessing on this dataset and store all information into database.
Phase 2: User enters the query for searching the any hospital location. Our system find the distance between the users enter query keyword and document location. Find the distance of keyword and document.
Display Result: This distance display on map and we recommend the place of the user.

IV. SYSTEM ANALYSIS

Following figure shows the breakdown structure of the system

T1: Database: Database construction.

T2: Data Query Index: System will get query and provide index of query.

T3: KD Graph Construction: System will construct directed weighted bipartite graph.

DOI: 10.15680/IJIRCCE.2018.0606040

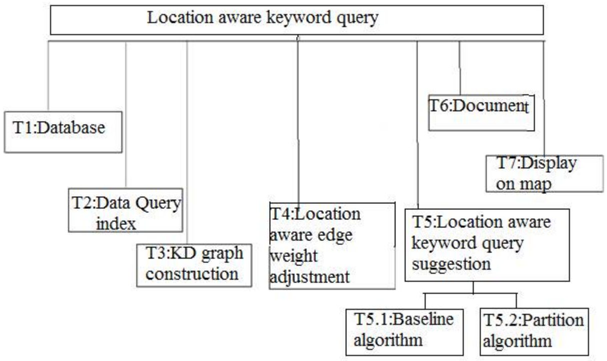


Fig. 2. Breakdown Structure

T4: Location Aware Edge Weight adjustment: Conform the edge weights in the KD graph built on the spatial connections the middle of the area of the query backer and the hubs of the KD-graph.

T5: Location-Aware Keyword Query Suggestion: Figure for every keyword queries a graph proximity score with admiration to kq, based on the random walk with restart procedure.

T5.1: Baseline Algorithm: A baseline algorithm (BA) used for location-aware suggestions. Steps of baseline algorithm are as follows:

1. To find the RWR based top-m inquiry recommend. 2.Start from one unit of active ink injected under hub Kq and the request in descending request.

3.Find the weight of every edge e is balanced based on q. 4.The algorithm returns the top-m candidate suggestions other than kq in C as the outcome.

T5.2: Partition Algorithm: A partition-based algorithm that isolates the keyword queries and the documents in the KD- graph G under aggregations. Steps of partition algorithm are as follows:

1. Toward every iteration, special case hub is processed, the active ink drops gradually and the end states met after too much cycle.
2. Provided for the vast number of iterations, the overhead of administering queue Q is huge.
3. Priority queue Q supports the partitions to be transformed in descending request of their keys.

Following table shows the result of the system:

4. For values of n from 500 on 10000, run a amount of examinations on randomly-ordered arrays size n and search the average number of comparisons for those examinations.
5. The algorithm returns the top-m candidate suggestions in C as the outcome.
6. Graph the average number of comparisons as a function of n Repeat the above things 14, utilizing an elective pivot selection strategy.

T6: Documents: Matched documents display to the user.

T7: Display On Map: Matched location display on map with distance.

Result of system are shown in following table:

Table 1 shows the number of specialist hospital added in the dataset. First column of table contain specialist such as Dentist,Audiologist,Gynecologist, Immunologist,Cardiologist, Orthopedic Surgeon, Ophthalmologist. There are 10 hospitals of specialist dentist,6 hospitals of specialist audiologist,9 hospitals of specialist gynecologist,8 hospitals of specialist immunologist,4 hospitals of specialist cardiologist,8 hospitals of orthopedic surgeon and 5 hospitals of specialist Ophthalmologist in the database.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

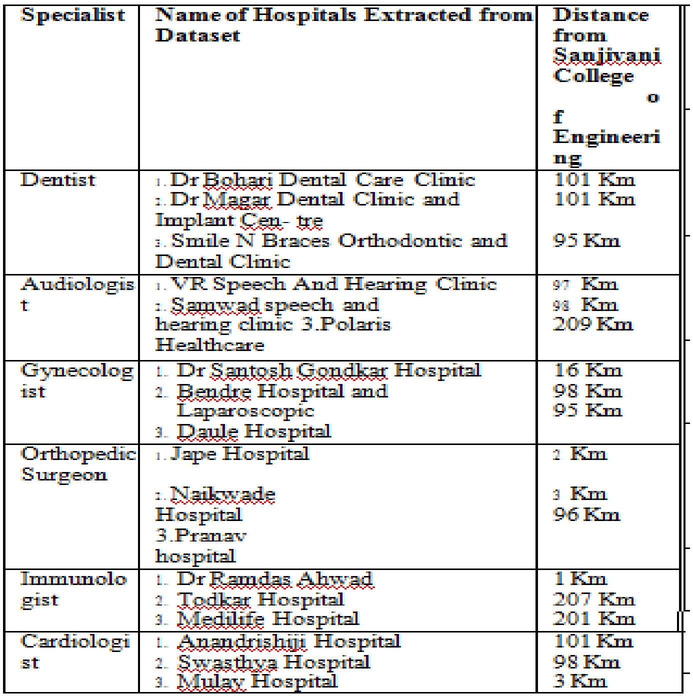
TABLE I
NUMBER OF HOSPITALS STORED IN DATABASE

Specialist	Number of Hospital in Database
Dentist	10
Audiologist	6
Gynecologist	9
Immunologist	8
Cardiologist	4
Orthopedic Surgeon	8
Ophthalmologist	5

Table 2 shows the result of system where current location of user is sanjivani college of engineering kopargaon. Dataset contain total 50 hospitals from different city for example kopargaon,nashik,pune,shirdi,yeola etc. Searching is based on specialist of hospital 7 categories of specialist hospital are included in dataset such as Dentist,Audiologist, Gynecologist, Immunologist,Cardiologist,Orthopedic Surgeon, Ophthalmologist. There are 10 hospitals of specialist dentist,6 hospitals of specialist audiologist,9 hospitals of specialist gynecologist,8 hospitals of specialist immunologist,4 hospitals of specialist cardiologist,8 hospitals of orthopedic surgeon and 5 hospitals of specialist Ophthalmologist in the database.

If user search is for spcialist Dentist and user location is sanjivani college of engineering. Dataset contain 10 dentist hospital from city pune,ahmednagar. system display the nearer hospital i.e hospitals of ahmednagar first is Dr Bohari Dental Care Clinic then calculate the distance from user current location i.e from kopargaon to ahmednagar which is 101 Km. In this way system extract the hospital according to query of user display the distance and route on map.

TABLE II
DISTANCE FROM CURRENT LOCATION TO EXTRACTED HOSPITAL



Specialist	Name of Hospitals Extracted from Dataset	Distance from Sanjivani College o f Engineering
Dentist	1. Dr Bohari Dental Care Clinic 2. Dr Magar Dental Clinic and Implant Cen- tre 3. Smile N Braces Orthodontic and Dental Clinic	101 Km 101 Km 95 Km
Audiologist	1. VR Speech And Hearing Clinic 2. Samwad speech and hearing clinic 3.Polaris Healthcare	97 Km 98 Km 209 Km
Gynecologist	1. Dr Santosh Gondkar Hospital 2. Bendre Hospital and Laparoscopic 3. Daule Hospital	16 Km 98 Km 95 Km
Orthopedic Surgeon	1. Jape Hospital 2. Naikwade Hospital 3.Pranav hospital	2 Km 3 Km 96 Km
Immunologist	1. Dr Ramdas Ahwad 2. Todkar Hospital 3. Medilife Hospital	1 Km 207 Km 201 Km
Cardiologist	1. Anandrishiji Hospital 2. Swasthya Hospital 3. Mulay Hospital	101 Km 98 Km 3 Km

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

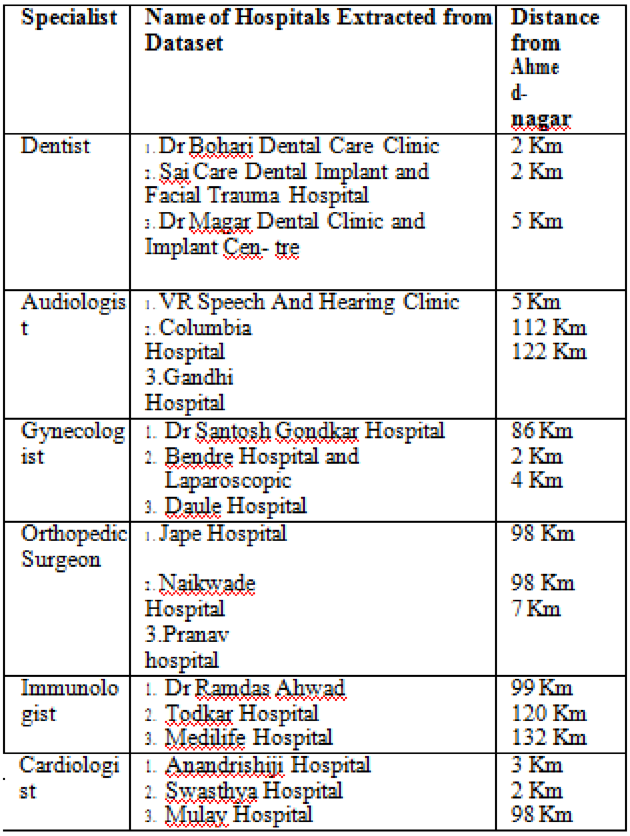
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

TABLE III
DISTANCE FROM CURRENT LOCATION TO EXTRACTED HOSPITAL



Specialist	Name of Hospitals Extracted from Dataset	Distance from Ahmed-nagar
Dentist	1. Dr Bohari Dental Care Clinic 2. Sai Care Dental Implant and Facial Trauma Hospital 3. Dr Magar Dental Clinic and Implant Cen- tre	2 Km 2 Km 5 Km
Audiologist	1. VR Speech And Hearing Clinic 2. Columbia Hospital 3.Gandhi Hospital	5 Km 112 Km 122 Km
Gynecologist	1. Dr Santosh Gondkar Hospital 2. Bendre Hospital and Laparoscopic 3. Daule Hospital	86 Km 2 Km 4 Km
Orthopedic Surgeon	1. Jape Hospital 2. Naikwade Hospital 3.Pranav hospital	98 Km 98 Km 7 Km
Immunologist	1. Dr Ramdas Ahwad 2. Todkar Hospital 3. Medilife Hospital	99 Km 120 Km 132 Km
Cardiologist	1. Anandrishiji Hospital 2. Swasthya Hospital 3. Mulay Hospital	3 Km 2 Km 98 Km

Table 3 shows the number of nearer hospitals extracted according to specialist of hospital. System shows the following result when current location of user is ahmednagar. There are three columns first column is specialist of hospital such as Dentist,Audiologist, Gynecologist, Immunologist,Cardiologist,Orthopedic Surgeon, Ophthalmologist. Second column is number of hospitals extracted when current location of user is ahmednagar. Third column shows current location of user. If user search is for specialist dentist then system extracts three hospitals such as 1.Dr Bohari Dental Care Clinic , 2.Sai Care Dental Implant and Facial Trauma Hospital ,3.Dr Magar Dental Clinic and Implant Centre. shows the distance and route on the map.

V. CONCLUSION

In this project,system provide a relevant user information need at the same time retrieve relevant document near to user loaction. For this evaluation user provide single keyword query then it calculate the distance based on query using partition algorithm. This system is related to application hospital so dataset contain hospital related information such as hospital name longitude and latitude of hospital,address and city of hospital. The search is based on specialist of hospital total six categories of specialist are stored in dataset namely Dentist,Audiologist,Gynecologist, Immunologist,Cardiologist, Orthopedic Surgeon. The dataset contain 50 hospitals. User can search specialist hospital then system extract three nearer hospitals from the user current location also system display the distance and route of hospital from current location.

ACKNOWLEDGEMENT

I would like to thank to my esteemed guide, Prof.P. N. Kalavadekar, whose interest and guidance helped me to complete the work successfully. Also under the valuable help of Dr .D. B. Kshirsagar(HOD Comp.Dept.) This experience will always steer me to my work perfectly and professionally and Prof.P. N. Kalavadekar (PG Co-ordinator) who has provided facilities to explore the subject with more enthusiasm.

REFERENCES

[1] Shuyao Qi, Dingming Wu, and Nikos Mamoulis, "Location Aware Keyword Query Suggestion based on Document Proximity",in IEEE,2015,pp.82-97.

[2] M. P. Kato, T. Sakai, and K. Tanaka, "A query suggestion log analysis", Inf. Retr., vol. 16, no. 6, pp. 725746, 2013.

[3] Bo Long, Jiang Bian, Olivier Chapelle, Ya Zhang, "Active Learning for Ranking through Expected Loss Optimization", in Information Science Vol. 220, 2013, pp. 269291.

[4] D. Wu, G. Cong, and C. S. Jensen, "A framework for efficient spatial web object retrieval", vol. 21, no. 6, pp. 797822, 2012.

[5] A. Anagnostopoulos, L. Becchetti, C. Castillo, and A. Gionis "An optimization framework for query recommendation, ", in Proc. ACM Int. Conf. Web Search Data Mining, 2010, pp. 161170.

[6] Y. Song and L.-W. He, "Optimal rare query suggestion with implicit user feedback ", in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 901910..

[7] Paolo Boldi, Francesco Bonchi, Carlos Castillo,"Query Suggestions Using Query-Flow Graphsrq', Knowledge Based System Vol. 82, 2015, pp. 2940.

[8] Q. Mei, D. Zhou, and K. Church "Query suggestion using hitting time", in Proc. 17th ACM Conf. Inf. Knowl. Manage, 2008, pp. 469478.

[9] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query recommendation using query logs in search engines", in Proc. Int. Conf. Current Trends Database Technol., 2004, pp. 588596.

[10] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log", in Knowledge-Based System Vol.73, 2015, pp. 311323.