A Hybrid Query Recommendation Technique in Information Retrieval

Neelanshi Wadhwa(✉), Rajesh Kumar Pateriya, and Sonika Shrivastava

Department of Computer Science and Engineering,
Maulana Azad National Institute of Technology, Bhopal, India
neelanshiwadhwa@gmail.com, pateriyark@gmail.com,
ms27ll04@gmail.com

Abstract. As the amount of information available online is enormous, search engines continue to be the best tools to find relevant and required information in the least amount of time. However, with this growth of internet, the number of pages indexed in search engines is also increasing rapidly. The major concern at present is no more having enough information or not; it is rather having too much information which is in numerous different formats, languages and without any measure of precision. Therefore, it is essential to devise techniques that can benefit the process of extracting useful information suitable for users' demands. Several mechanisms have been developed and some methods have been enhanced by researchers from all over the world to generate better or more relevant query that can be provided as suggestion to the user for enriched Information Retrieval. The objective of this paper is to summarize and analyze the various techniques adopted to optimize the Web Search process to support the user. The existing strategies developed in this scenario are also compared using standard IR metrics to evaluate the relevance of results.

Keywords: Query recommendation · Query logs · Information retrieval

1 Introduction

1.1 Fundamental Information Retrieval Process

The basic Information Retrieval process in a search engine is depicted as:

- Data is available in form of documents or Webpages. These are organized in a specific format and an index is created.
- As the query is fired to the search engine, best matching entries from the index are selected.
- Simultaneously, a learning algorithm is developed that guides the rank updation of results.
- Finally, the results are displayed to the user in descending order of ranks.

This process is displayed in Fig. 1.

P. K. Singh et al. (Eds.): FTNCT 2018, CCIS 958, pp. 165–175, 2019.
https://doi.org/10.1007/978-981-13-3804-5_13

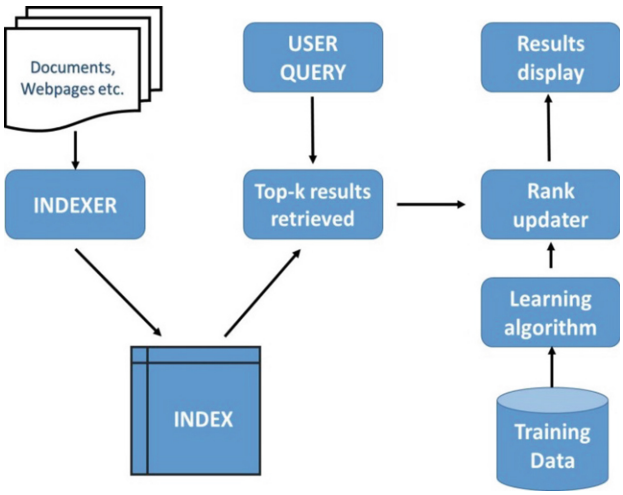

Fig. 1. Basic IR process

1.2 Motivation Behind Query Recommendation

With such huge volume of information available, it is a challenging task to find relevant information that meets the needs of the user using simple search queries. This problem arises mainly because the queries submitted by the user to the search engine are usually very short and turn out to be ambiguous. These queries fail to convey the user's requirements precisely. Consequently, lots of results retrieved by the search engine might be irrelevant with respect to the users' needs due to the implicit ambiguity in queries. Also, most of the users go through only the first one or two pages of the retrieved results, there is a possibility that they would fail to notice many relevant results obtained in successive pages that are left unread. Hence, Web Search becomes an iterative trial and error process by changing queries each time so as to fulfill the information need. Alternatively, the users may not want to reformulate their queries in order to avoid the additional efforts while searching.

The compulsive need to extract useful piece of information for the user from such voluminous store of web pages leads to various issues such as the ranked list problem, ambiguous query problem, obtaining results semantically dissimilar to what the user desires and so on.

To overcome all the above discussed limitations of search engines, research is directed towards generating clear and more efficient queries as suggestions to users. These newly generated queries tend to retrieve more relevant results as per the users' requirements. There are number of strategies proposed in this direction, each of them involves a unique yet effective method of query recommendation. Some of these strategies exploit various kinds of background knowledge or data available to study users' behavior in order to yield focused results.

1.3 Terminology

Query recommendation can be defined as the process of reconstructing queries entered by the user so as to provide the user with better results which are more relatable to user context or domain.

Query Log – The search engine records an entry for each user (each query) which contains entries such as

- Query q given by the user
- User or session ID
- The URLs which the users clicked from the displayed results
- Rank of the web page accessed etc.

The rest of the paper is organized as follows: Sect. 2 gives a detailed review of related work in query recommendation and various proposed models. A comparison chart based on advantages, shortcomings and performance metrics has also been described. Section 3 describes a unique strategy proposed to increase the efficiency of existing query suggestion process. In the end, there is the conclusion and a brief discussion of future work.

2 Literature Survey

Most approaches of query recommendation focus on users' previously submitted queries which are analogous to the current query either in terms of content or in click context. Liu et al. [2] propose a framework very different from its preceding methods, which attempts to obtain user's information need by means of click-through logs as shown in Fig. 2. It is observed that although the clicked documents may not always be relevant to user but the snippets that make the users click on the links better represent users' need. Therefore, two snippet click models are built according to this finding and corresponding algorithms are described. These methods do enhance the query recommendation process, but the users' context (or background) cannot be ignored as it provides a good insight to understand the domain of the query. Hence, users' context features should be added to create a universal model.

In paper [3], author aims on an eminent problem of Web searches, known as the ranked list problem, which has emerged as a consequence of two major activities; the first being the fact that search engines mostly present the results in the form of ranked lists, and the second being the semantic ambiguity of users' queries put in to the search engines. Search engines usually retrieve and rank documents irrespective of the probable different semantics of query terms due to short and ambiguous queries. Moreover, since it is a general notion amongst most users to explore only the few pages of ranked results, the possibility that users might miss many relevant documents retrieved in subsequent and unread pages is quite huge. To overcome these weaknesses, users perform iterative cycles based on hit and trial to reformulate their query. In the attempt to capture new and more relevant documents in the top p (assumed) ranked positions, users submit minor variants of the original query at each step which may express their needs in a better way. However, this behavior leads to the retrieval of almost the same documents in the first

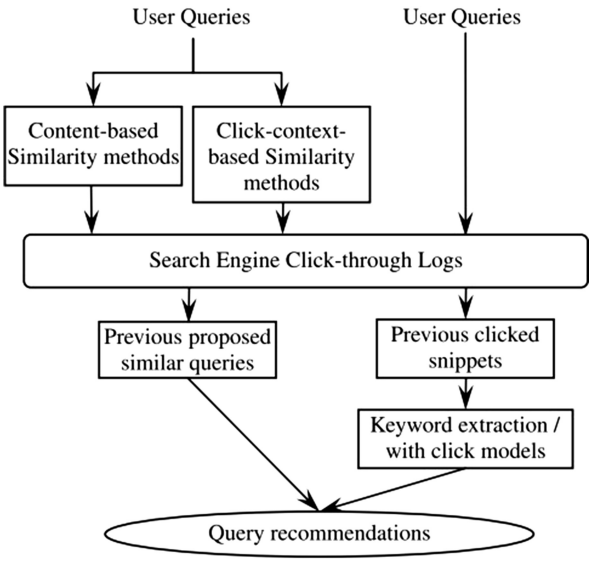

Fig. 2. Snippet click model [2]

positions of the ranked list, thereby vanishing the efforts of users. They propose an iterative query disambiguation procedure that attempts to create and suggest disambiguated queries, which potentially may retrieve both new and relevant documents [3].

Baeza-Yates et al. [5] propose a technique to suggest a list of queries related to the query that is submitted. These new queries are can be used by the users' to redirect the search process. Their method consists of a query clustering process which identifies queries that are similar in meaning or definition and groups them to form a cluster. The clustering procedure is based on the historical activities of users recorded in the query log of the search engine. After finding a set of related queries, they are ranked as per certain relevance criteria. The two measures used in this method to calculate rank of a query are:

1. Similarity – That defines how much the new query is alike to the input query. This is calculated using term-weight vector quantities for each query.
2. Support – That is the fraction of the documents returned by the query that are clicked by the user. This is obtained using query logs.

Although this technique yields good results in terms of precision of suggested queries as compared to the original query, but the overall performance of this technique may not prove to be good enough for huge number of users and queries with corresponding query logs. Also, there might be sets of similar queries that have common words but do not have same set of clicked URLs (or vice versa), which is not captured in this technique.

Bordogna et al. address the ranked list problem that arises in internet search whenever a user puts in a very short request or query. It is because the entered words generate ambiguity and convey many different meanings at the same time pertaining to many different scenarios. Hence they have devised an iterative mechanism for disambiguation of queries that contains following major steps and depicted in Fig. 3:

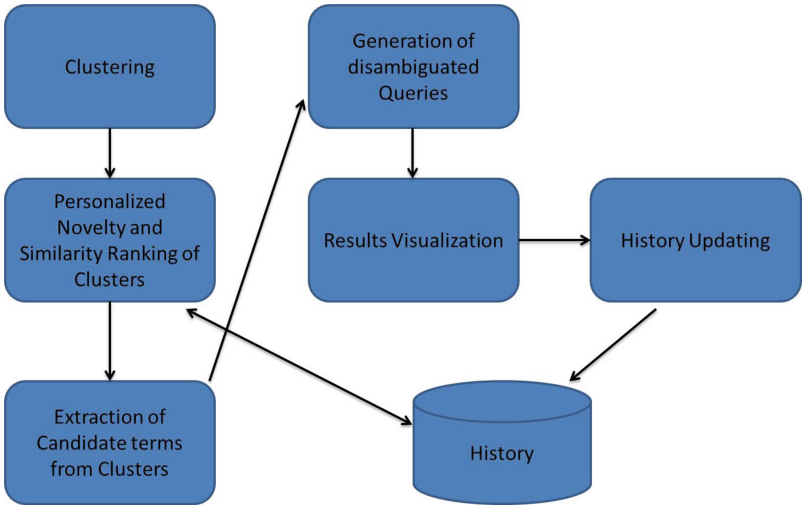

Fig. 3. Query disambiguation scheme [3]

1. Clustering – This is the very first step where the results obtained from a web search made by a user are grouped together into different clusters.
2. Personalized novelty and similarity ranking – In this step, each cluster is assigned a score which is calculated from two parameters, similarity of cluster content to query and originality of cluster with respect to user's past behavior.
3. Extraction of cluster candidate terms – From each cluster, a certain no of set of weighted terms along with their weights is selected. These will be used further in creating new query.
4. Disambiguated query creation – Now, first n number of highest weighted candidate terms are used to form the new query.
5. Displays of results – The clusters are shown according to the ranking and one disambiguated query for every cluster is also shown to the user.
6. History updating – Finally, when user selects one of the new queries, that query is taken as more relevant to repeat the entire process until the user stops submitting further queries [3].

This technique is unique and would considerably achieve good results in desired time, but the idea of displaying clusters to the users is that appropriate as the user is only interested in getting the required information in form of documents or web pages. Also, the process is time consuming as it contains complex computations at each step of each iteration.

Zhu et al. [8] claim that instead of just looking for higher relevance value of queries, it is more advantageous to directly recommend queries with greater utility. They define query utility as "The information gain that a user can get from the search results of the query as per the user's original search objective." For this a Query Utility Model (QUM) has been proposed by them to obtain query utility. This parameter, Query Utility, is calculated by studying users' query reformulation and their click activity in a particular search process. Query utility is said to contain two components:

1. Perceived utility – This is based on the attractiveness of the search results displayed for a particular query, which further increases the chance that a URL will be accessed by the user.
2. Posterior utility – This is the satisfaction that a user obtains as information gain by clicking any search result of the query.

Further, a unique dynamic Bayesian network is developed to compute Query Utility based on above discussed parameters, known as Query Utility Model (QUM). A publicly released query log is used to test the performance and compare it with other already renowned methods available for query recommendation and experimental results are found very promising. The evaluation is done on below two metrics as described in Eqs. (1) and (2):

1. Query Relevant Ratio (QRR)

$$\mathrm{QRR}\,(q) = \frac{\mathrm{RQ}(q)}{\mathrm{N}\,(q)} \tag{1}$$

Where
RQ(q) is the total frequency of query q with relevant results clicked by users,
And N(q) is the total frequency of query q issued by users.

2. Mean Relevant Document (MRD)

$$\mathrm{MRD}\,(q) = \frac{\mathrm{RD}\,(q)}{\mathrm{N}\,(q)} \tag{2}$$

Where
RD(q) is the total frequency of relevant results clicked by users when they use query q for their search tasks,
and N(q) is the total frequency of query q issued by users.

Many other different ways and models have been proposed to improve the process of query recommendation. He et al. [6] focus more on the ordering of queries within a search session, as they have high degree of correlation. These queries should be analyzed sequentially so as to know the information need of the user in a better way. However, to set up this kind of model on huge data set (i.e. query log of all users available) might require training the data. Nguyen et al. [7] present a different view to some extent. They say that Web-page recommendation can be more efficient by integrating the Web usage knowledge as well as the domain knowledge related to the key concepts of the query. They have proposed models to establish relation between the two and use them to provide better recommendation to the user.

Song et al. [1] propose to integrate all essential features that enhance the query recommendation process by creating a hybrid recommendation strategy. To begin with, a concept extraction method is efficiently modified to find queries similar in terms of concept. These concepts are mined using the web-snippets of the original query. To better represent co-related queries, a bipartite graph (between query and concept) is created that helps in finding similarity. Secondly, it is also observed that many times

URLs contain significant tokens that may depict the actual webpage contents. URLs are separated into tokens using the TF-IQF model. Further, three vital similarity features are exploited to develop a hybrid sematic similarity model for query recommendation. These aspects are defined below:

1. Clicked document – The set of URLs clicked or set of documents accessed for a particular query can be assumed to have alike content in terms of concept.
2. Associated Query – It is natural that queries that contain more number of identical terms (with respect to meaning i.e. synonyms), they will have a higher value of similarity.
3. Reverse Query – Two or more queries that cause the users to click on the same URL are obviously similar and relevant queries. They are called as reverse queries.

Finally, the unique approach suggested takes into account all the above measures in weighted form as any one of them is not sufficient. This hybrid approach is formulated as:

$$\mathrm{sim}(p,q) = \alpha * \mathrm{sim}_{\mathrm{doc}}(p,q) + \beta * \mathrm{sim}_{\mathrm{ass}}(p,q) + \gamma * \mathrm{sim}_{\mathrm{rev}}(p,q)$$

Where
$\mathrm{sim}_{\mathrm{doc}}(p,q)$ is the clicked document similarity,
$\mathrm{sim}_{\mathrm{ass}}(p,q)$ is the associated query similarity
And $\mathrm{sim}_{\mathrm{rev}}(p,q)$ is the reverse query similarity.
α, β, and Υ are three real constants between 0 and 1, and they satisfy the restriction of $\alpha + \beta + \Upsilon = 1$.

Now, to find the optimal weights of these real constants, experiments are done by randomly adjusting their values. The results are evaluated using the standard IR metrics Precision, Recall and F-Measure. It has been shown via experimental analysis that the hybrid scheme achieves better Precision and Recall simultaneously as compared to applying each individual similarity measure separately.

Table 1 shows the comparison of various query recommendation methods.

Table 1. Comparative analysis of existing techniques

Author	Problem addressed	Technique	Advantage	Disadvantage	Metrics
Liu et al. [2]	Getting users' exact information need	Snippet click model	More efficient than current practical search engines	Users' prior search context not taken into account	No standard metrics used for performance evaluation
Gloria et al. [3]	Ranked List problem	Query disambiguation	Retrieves new documents	Unnecessary display of cluster along with best query suggested	No standard metrics used for performance evaluation
Zhu et al. [8]	Reformulating Queries	Query Utility Model (Bayesian Network)	Achieves useful recommendation of queries	Automated procedure not formed	Query Relevant Ratio (QRR)
Song et al. [1]	Ambiguous and Short Queries	Hybrid query similarity model	Considers all major similarity measures and gives good results	Purely experimental method of parameter evaluation	Precision, Recall, F-Measure

3 Proposed Model

In this section, we propose a collaborative query recommendation model which incorporates all major similarity measures in a weighted scheme to generate a resultant similarity score for queries. This resultant score will be the deciding factor to select the best query (or set of queries) to be provided as a suggestion to the users so as to narrow down their search process and retrieve the most relevant information.

In the proposed scheme we attempt to calculate the weight of different similarity measures by using Genetic Algorithm as the learning algorithm. Initially random weights can be assigned to create a set of chromosomes and then on applying genetic algorithm to these chromosomes, we can obtain optimal values. The set of weights that would yield the highest similarity score can be taken as the best chromosome or the best solution to the problem. Further, we will generate an alternate query based on achieved similarity measure and the optimal weights. The set of queries obtained can be provided to the user as suggestions. The working flowchart of the proposed methodology is given below in Fig. 4 and the detail working of Genetic algorithm is described in Fig. 5.

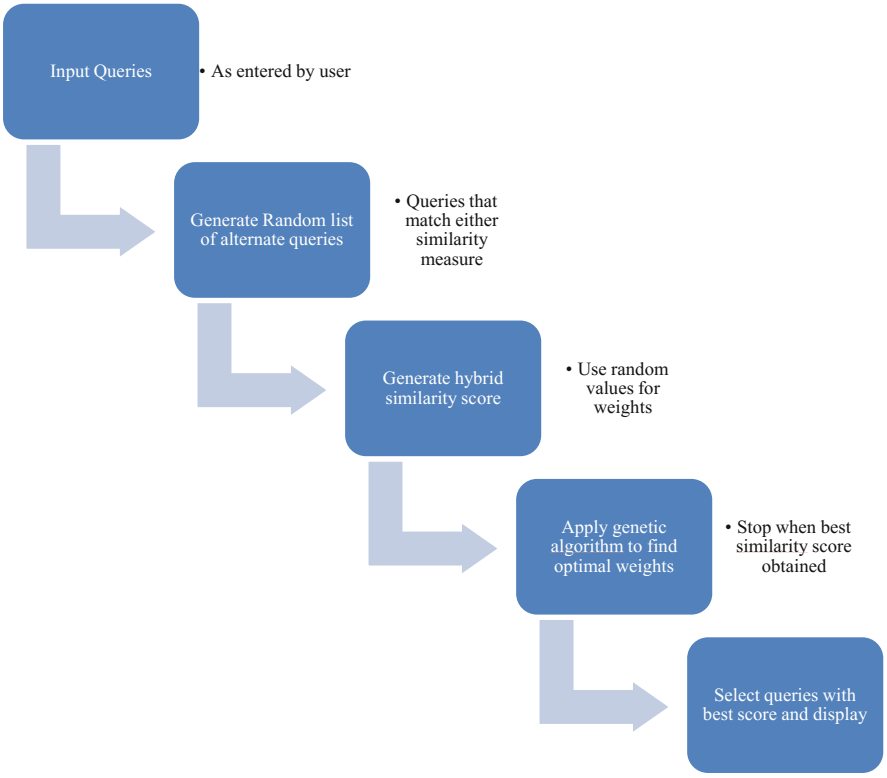

Fig. 4. Process flow for proposed scheme

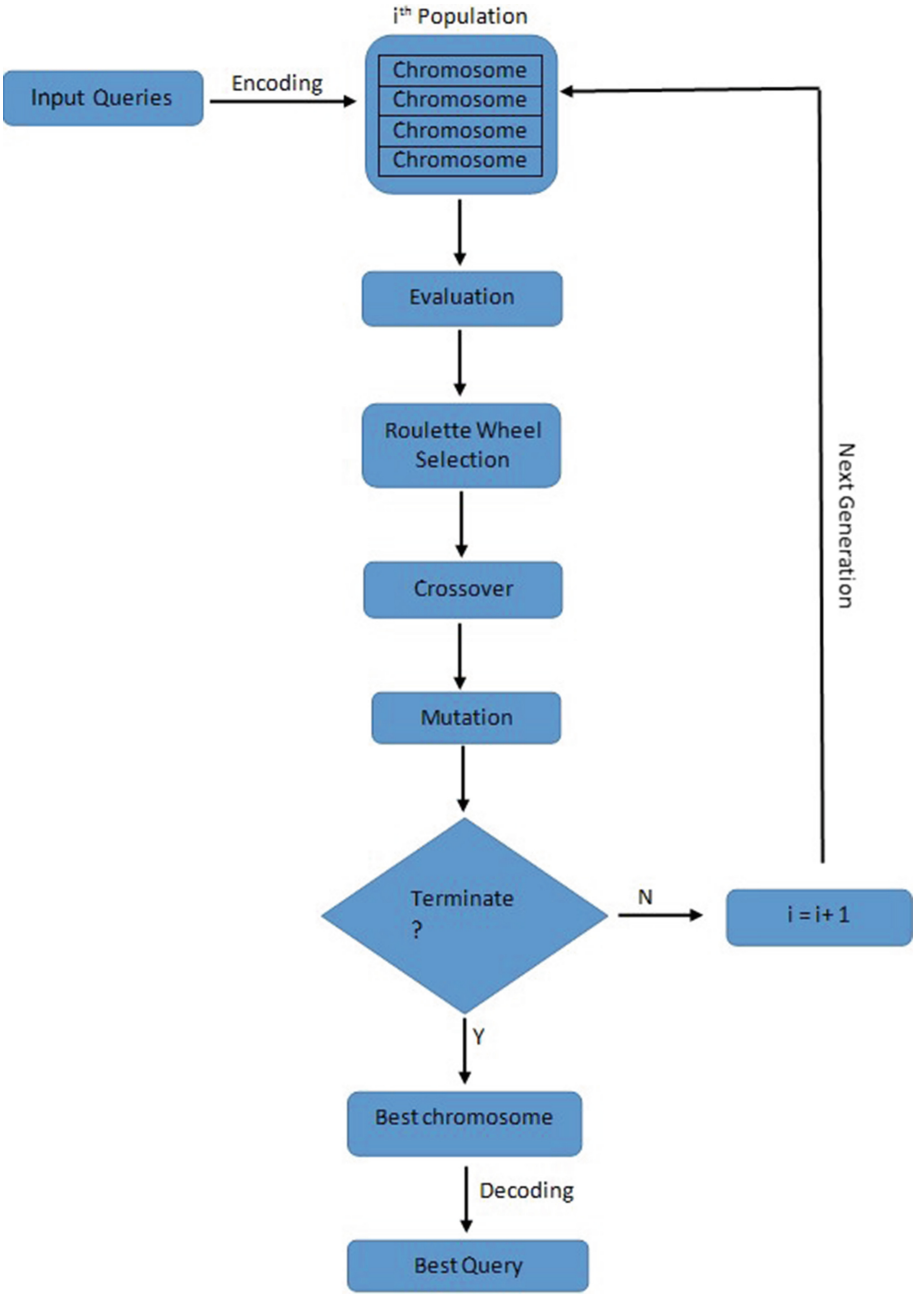

Fig. 5. Proposed application of genetic algorithm

4 Conclusion

In this paper, we have described the Query recommendation strategies that can be applied to improve the performance of search engines. We have also discussed what issues led to the idea of Query reformulation/Recommendation. Query recommendation has now become an inseparable part of search process as it saves users' time involved in rephrasing queries and also helps users satisfy their need for information. This process is also beneficial when a naïve user enters an ambiguous query. At the same time, it guides the users towards their ultimate information goal even when they themselves are unable to clearly express their requirements due to lack of knowledge or experience.

Thereafter, we have proposed a hybrid comprehensive strategy for effective query recommendation that is based on learning algorithm. By making use of Genetic algorithm significant results can be achieved for optimized values of the weights of similarity measures. This would lead to the generation of better and more relevant and unambiguous queries to be recommended to the user.

References

1. Song, W., Liang, J.Z., Cao, X.L., Park, S.C.: An effective query recommendation approach using semantic strategies for intelligent information retrieval. Expert Syst. Appl. **41**, 366–372 (2014)
2. Liu, Y., Miao, J., Zhang, M., Ma, S., Ru, L.: How do users describe their information need: query recommendation based on snippet click model. Expert Syst. Appl. **38**, 13847–13856 (2011)
3. Bordogna, G., Campi, A., Psaila, G., Ronchi, S.: Disambiguated query suggestions and personalized content-similarity and novelty ranking of clustered results to optimize web searches. Inf. Process. Manag. **48**, 419–437 (2012)
4. Zahera, H.M., El Haddy, G.F., Keshk, A.E.: Optimizing Search Engine Result using an Intelligent Model (2012)
5. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query Recommendation Using Query Logs in Search Engines. In: Lindner, W., Mesiti, M., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) Current Trends in Database Technology - EDBT 2004 Workshops. LNCS, vol. 3268, pp. 588–596. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30192-9_58
6. He, Q.: Web query recommendation via sequential query prediction. In: IEEE International Conference on Data Engineering, 1084–4627/09 (2009)
7. Nguyen, T.T.S., Lu, H.Y., Lu, J.: Web-page recommendation based on web usage and domain knowledge. IEEE Trans. Knowl. Data Eng. **26**(10), 2574–2587 (2014)
8. Zhu, X., Guo, J., Cheng, X., Lan, Y.: More than relevance: high utility query recommendation by mining users' search behaviors, In: CIKM 2012, 29 October–2 November 2012, Maui, HI, USA (2012)
9. Habibia, M., Mahdabib, P., Popescu-Belis, A.: Question answering in conversations: query refinement using contextual and semantic information. Data Knowl. Eng. **106**, 38–51 (2016)
10. Shanna, A.K., Aggarwal, N., Duhan, N., Gupta, R.: Web search result optimization by mining the search engine query logs. In: International Conference on Methods and Models in Computer Science (2010)

11. Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A.: An optimization framework for query recommendation. In: WSDM, pp. 161–170 (2010)
12. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: SIGKDD, pp. 407–416 (2000)
13. Yadav, U., Duhan, N., Kaushik, B.: Relevant page retrieval and query recommendation using semantic analysis of queries. Int. J. Sci. Eng. Res. **4**(7), 694 (2013)
14. Deepak, G., Priyadarshini, J.S., Hareesh Babu, M.S.: A differential semantic algorithm for query relevant web page recommendation. In: IEEE International Conference on Advances in Computer Applications (ICACA) (2016)
15. Sahu, S.K., Mahapatra, D.P., Balabantaray, R.C.: Analytical study on intelligent information retrieval system using semantic network. In: ICCCA (2016)