3. EXPLORATIVE APPROACH OF THE PROPOSED RESEARCH METHODOLOGY

3.1 GLIMPSES

The inconceivable evolutionary dynamism of the web increases the scope of search engines in the process of IR. At maximum instances, the query keywords are either assumptions or non-existent coinages of dictionaries to procure the anticipated results. This happens due to the very limited background knowledge about the domain of the users who attempt to retrieve exact and relevant results in their search. In such cases, nothing except query recommendation techniques can prevent the search engines from mustering totally irrelevant and redundant results for the search. Precisely, query recommendation sustains its impeccable support through making appropriate suggestions to replace the inappropriate queries fed by the users to the search engines. Therefore, the user is privileged with better search results by utilising the recommended queries as a desirable replacement to their initial query. The explicit framework of this research aspring to enhance the exisiting methodologies of web search is precisely presented in this chapter.

3.2 FRAMEWORK OF SEARCH QUERY RECOMMENDATIONS IN WEB IR

The navigational history of the web users and the clues for their respective search pattern are generally available in the query log files. An empirical exploration of these data is expected to add a desirable twist to the users’ search through the supplementation of better queries suggested as an outcome of this research. The framework for providing such suitable query recommendations is discussed in this chapter. Figure 3.1 obviously, presents the sketch of the entire research carried out in four phases.

This tetra-faceted research focuses on time confined search query recommendations using query log and user profile. While Phase I of this research analyses the query logs to identify the users and their corresponding behavior patterns, Phase II focuses on the query recommendations using query logs without time constraint. Phase III exhibits time weight as another significant parameter of this research and re-ranking the recommended queries. Phase IV consolidates the evaluations made on the recommended queries.
3.3 METHODOLOGY

This research presents an approach to overcome the difficulties in web IR. The significant task of the research is explained in the subsequent sections.
3.3.1 Innovative Strategies of Query Log Analysis

Generally, the query log files of the search engines are deposited with enormous information through each individual access of the users in search of specific details. The storage of this minuscule details constituted in the query log files assist in extracting very precise and relevant data requisite from the web source. In fact, mining query log files turns to be the best strategy to explore the search target of the users, their individual sessions, apart from their individual intent grasped through a set pattern of their search. The knowledge about the above stated fields of each search comprehended through the query logs, perhaps, helps to improvise quality of answer lists besides, providing a set of recommendations to the user instead of using the initial query. The noisy and irrelevant data from the log files are eliminated and it is represented in a Common Log format (CLF). At this point, the details about the distinctive and inimitable web users are retrieved from the query log files. Consecutively, the classification of each search session along with its respective users is classified. Further, an exclusive examination of the cleaned pre-processed query log files are clubbed with its corresponding access pattern. Then the access pattern against each query submitted by the user is generated. From these frequent patterns, the URL click count is calculated through trivaricated techniques namely, Sequential URL Click Count, Bit Vector URL Click Count and Hash based URL Click Count. Here, hash based technique proves to be a noteworthy effort with respect to memory and speed. To state precisely, the application of hash based techniques in query recommendation is edified in the consecutive phases of the research.

3.3.2 Time Invariant Query Recommendations from Search Engine Query Logs

The futile procuring of redundant and irrelevant results of a search engine happens because of the inadequacies that exist in the comprehending strategies fed to the search engines to understand the intension of the user. Query recommendations has become the vital indispensable web search activity that kindles the user to frame the meaningful queries which are used to retrieve the relevant and accurate results. Perhaps, the recommendations made are based on the real intent of the user, comprehended through the study pursued from the past search histories. Above all these recommendations to supplement the query input are provided by considering the query and its navigations occurred at any time period. As the time component is exclusively used for session identification, it does not become a constraint.
The initial phase of this study focuses on mining the query log files with an aspiration to determine the similarity between the query keywords and URLs. For instance, when the URL \( U \) is clicked for a couple of queries namely, \( Q_i \) and \( Q_j \) then, the queries are perceived to be similar ones as they cling on to same URL. The second phase of the research exhibits the process of creating both query and URL clusters promoted by combined similarity measure generated in the first phase. In fact, these clusters assume a vital role in making recommendations to assist the user to frame their futuristic queries. Moreover, this combined similarity based approach provides the recommendations to the user about the URL selection. The recommendations are given to the naive users by using the query keywords from the expert’s, besides that both the users are having similar search intent. This technique is called as collaborative recommendations. Other than the query keyword and URL, the concepts available in the clicked web snippets are also considered to provide the recommendations.

Generally, the search results from top to the bottom of resultant web snippets are scanned by the users who instantaneously makes discretion to click a specific snippet as it makes an appeal to be an essential requisite. While the concepts in the clicked documents are addressed as concepts with positive preferences, the concepts in the un-clicked documents are titled as concepts with negative preferences. At this juncture, the users are clustered based on both positive and negative concept preferences. The users observed to have exhibited similar search intensions are then clustered with a purpose to add efficacy to the recommendation process.

### 3.3.3 Time Heuristics Query Recommendations and Re-ranking of Recommended Queries

The recommendation technique is based on the weight assigned to the queries. The queries occurred successively at a given time period are clustered based on the hub, authority and time measure. To start with, the query keywords are pre-processed as a preliminary effort. Then, the frequently accessed queries and URLs are identified from the query log using the modified frequent pattern generation algorithms PrefixSpan and UDDAG. Next hub and authority weights are calculated for the frequent items. The similar queries occurred at a given time period are clustered and the weight t-measure is assigned to the queries in the cluster. This is performed with an intention that the similar queries occurred more recently have higher weight than the earlier queries. The similar queries are clustered based on the temporal characteristics of historical click-through data. The intuition is to reveal that more
accurate semantic similarity of queries can be obtained by considering the timestamps of the log data. The cluster generated in this approach is used to provide query and URL recommendations to the user. Finally the method has been evaluated using real data set of AOL search engine query log.

The search user \( U \) is classified corresponding to the query \( Q \) with exclusive reference to the below stated denominations.

- The User \( U \) and the query \( Q \) are fresh entries
- The User \( U \) is a new entry whereas the query \( Q \) already exists
- The User \( U \) exists in the list whereas the query \( Q \) is a fresh entry
- The User \( U \) and the query \( Q \) already exists in the query log.

In such instances the recommendations shall be given in three different ways namely:

- Identification combined with favourite query of the user.
- Clustering of those users who possess similar interest and recommending such users tracked and inferred from the access logs.
- Clustering of similar queries with a view to identify the favourite query of the cluster before making recommendations.

Consecutively, the queries recommended are re-ranked based on the preferences and the t-measure of the queries. The correlation between the ranking orders is computed. The ranking techniques are evaluated by using the existing measures namely, Euclidean Distance, Manhattan Distance, Area under Curve (AUC), Ordered Area Under Curve (OAUC) and Accuracy. The ranked queries are evaluated using Mean Reciprocal Rank (MRR) measure. Finally the method has been evaluated using a query log of the search engine.

3.3.4 Personalized Query Recommendations using Concept based Hybrid User Profile

The concept based hybrid user profile is generated from the user log, query log and concepts in the click-through snippets to re-conduct the recommendation generation phase. The concept based hybrid user profile is used for recommending and re-ranking the queries. The given technique is very efficient and scalable; it is particularly effective in generating suggestions for rare and newly occurring queries. The major contributions for recommending the queries in this phase is stated subsequently:
• User’s preferences in their navigations against the query is analysed from the user log. Here, the information about the user is recorded implicitly.

• The past queries and search behavior is analysed from the user’s implicit feedback and the user’s concept intent is retrieved from the query log.

• The concepts are retrieved from the clicked web snippets and they are pre-processed. The relationships between the concepts are represented as a similarity matrix.

• The queries are recommended using User log, Query log and Concept log.

The recommended queries are re-ranked using the favourite query and URL click count. The recommended queries are evaluated using kappa measure and coefficient of variance.

3.4 SUMMARY

The methodologies proposed for the time-confined query recommendations organised in this tetra-faceted research are elaborately discussed with adequate illustrations in the consecutive chapters. The techniques proposed are experimented with the existing query log files of a search engine. Besides, it is also edifies that the proposed technique may be used in any kind of IR process that depends on the analysis of the log files.