1. INTRODUCTION

1.1 GLIMPSES

"It is often necessary to make choices with sufficient personal experience of the alternatives. In everyday life, we rely on recommendations from other people either by word of mouth, recommendation letters, movie and book reviews printed in newspapers, or general survey" - Paul Resnick.

The massive utility of web resources in recent scenario has turned to be an indispensable commitment for numerous reasons. Clinging on to the web information from a microcosmic level to the macrocosmic level has been growing over the last three decades. At the same time, the inconceivable boom of information available in the websites simultaneously throws the challenge of retrieving the precise and appropriate information at the time of need. To state the precise statistics of active websites, the March 2012 survey of Netcraft that figures around 644,275,754 websites may be quoted. This survey aids to comprehend how the web appears to be a panacea due to its inevitable applications in several facets of life.

Moreover, the web information is the mostly sought after powerful platform for working, studying, searching information, besides, being in touch with our friends. Apparently, the unpredictable amount of web information available becomes a menace of experiencing ambiguity in the web search. To prevent the web users from getting overwhelmed by the quantity of information available in the web, several strategies are proposed. These strategies attempt to solve the tedious information exploration process of the user, through Information System, Information Filtering and Recommendation Systems.

1.2 EXPLORING INFORMATION

Generally speaking, a web exploration significantly relies on the essential web applications like through Information System, Information Filtering and Recommendation Systems [HSS01].

Information System (IS) is a process which sustains its support to extract all possible items namely text, video, image, audio etc.

Information Filtering (IF) is an approach which strategically supplies precise and apposite data to those who search for it. The functioning capacity of this process involves a desirable
working condition on hefty data with the sustenance of the demographic details obtained through the profiles of the web users besides, their interests. While in the filtering process, main goal is the removal of data from an incoming stream, but the retrieval process focuses on finding relevant data in a stream.

Recommendation Systems (RECSYS) prove itself to have a pivotal function in the recovery of data from the web. This is addressed as web Information Retrieval (IR). The purpose of a recommender system is to observe and fix the user’s genuine search intention and to supply the requisite appropriately. A recommender system generally evaluates the user’s profile to certain reference features and foresees the rating that a user would give to an item.

Even with a little or hardly any information the user gains the privilege of information retrieval through this system [RV97]. Significantly, this system also advocates alternate suggestions in connection to the genuine intention of the user in a search. This actually helps the users to opt for the quick and best results. RECSYS aspire to procure interesting items besides combining them to the stream of information, whereas IS focuses on elimination of items from the stream. Burke [B02] defined that the RECSYS are tools that have the effect of guiding the user in a personalised way to interesting or useful objects in a large space of possible options. G. Adomavicius et al. [AT05] defined that RECSYS is a set of users $U$, a set of items $I$, and a utility function $f: U \times I \rightarrow R$, where $R$ is a totally ordered set of recommendations. With respect to Burke’s works [B02] [B07] the conventional recommendation approaches are classified into five types namely, collaborative, demographic, content-based, case-based and hybrid recommendation.

Collaborative Recommendation: This "people-to-people" [SKK01] recommendation technique, functions on the preferences deciphered through the similarity of search behavior of the web users. This may be illustrated with the user $U$ and the query $Q$. Here, when the user $U$ triggers a query $Q$, the collaborative filtering system supplies recommendation on the query $Q$ in a couple of steps. Primarily, this system attempts to explore a set of users $U$ with similar interests with respect to $Q$. Then it musters the queries issued by the users $U$. Collaborative approaches may also be called as customer-to-customer correlation as they have their firm roots on the perception of other customers. In fact, this similarity may be utilized in recommending the products to their unique customers. Collaborative filtering differentiates the customers’ neighbours easily from the history of information.
**Demographic Recommendation:** The objective of the demographic RECSYS is to categorize people with respect to their personal demographic details [B02] [B07], in addition to making suggestions corresponding to the users’ demographic profiles. Here, the assumption is that the users within same demographic class would possess similar interests [RRS11]. Customizing the suggestions is pursued with the precocious references like the users’ age, nativity, linguistic proficiency, and so on.

**Content-based Recommendation:** Content-based approach is absolutely dependent on the search histories and navigational behavior of the user. A content-based approach chooses data items depending on the correlation between the content of the data items and the user’s preferences as opposed. The goal of content-based RECSYS is to process the content of the items and to recommend items based on user’s profile.

**Case-based Recommendation:** Case-based reasoning [LRT05] extremely depends on the computational model that aspires to decipher the search intension with specific reference to the previous problem solving experiences. The solution retrieval of a query is automatically referred at the submission of each new query. Each case experience gets stored in a data-store for futuristic application. A case model is an entry in the data-store with problem and solution descriptions.

**Hybrid Recommendation:** Hybrid methods combine multiple recommendation techniques to compensate for limitations of single recommendation technique.

The recommendation strategy elaborated in this research, advises the utility of the previous queries already stored as a web resource and exist in the search histories. A thorough analysis of this previously available search information relevant to one’s query, besides, an overt comprehension of the search behavioral pattern of the previous users through the clicks made may be studied from the query log files. The recommendation technique proposed here is actually a hybrid approach, which is a combination of content based and collaborative strategies. A combination of these dual strategies may be suggested for other IR systems also. The consecutive sections brief on web mining, process of web usage mining, problem identification, scope of the research, objectives of the research and organization of the thesis.
1.3 WEB MINING

Web mining is one of the data mining techniques used to automatically discover and extract the interesting and potentially useful patterns and implicit information from web documents and services [E96]. Exploring and extracting precisely pragmatic knowledge from web data is called web mining. Web mining is indispensable to enhance the utility of web. Application of data mining techniques in the World Wide Web (WWW) is called as web mining [SPV05]. Web mining is used in four significant fields namely, Resource finding, Information election and Pre-processing, Generalization and Analysis.

*Resource finding:* The task of retrieving the anticipated web resource through exploration.

*Information selection and Pre-processing:* The process of making automatic choices while pre-processing to obtain a definite data from the retrieved web resources.

*Generalization:* It is an automatic method to examine general patterns at individual web sites as well as across multiple sites.

*Analysis:* It is a method of validation and/or interpretation of the mined patterns to reinstate the quality of results observed.

1.4 WEB MINING AND ITS CLASSIFICATION

Categorically stating, web mining exhibits partitions namely, *web content mining*, *web structure mining* and *web usage mining* based on type of data used for mining processes. However the prevalent types are web content mining and web usage mining. In one, web structure is considered as part of web content; while in the other, web usage is considered as part of web structure. The main goal of all the three classification types is the method of knowledge discovery of inherent, unidentified and potentially valuable information from the web. Each of them focuses on varied mining objects of the web. Figure 1.1 describes the web categories and their objects. Subsequently, a brief introduction is given for each of the categories.

*Web Content Mining:* It is the process of ascertaining information or resource from millions of sources across the WWW. Web content mining will perhaps be distinguished from two perspectives namely, the agent-based technique and the database approach. The first approach aims at enhancing the data finding and filtering. The second technique aspires for modelling of the data available on the web into more controlled forms with the intention of
Figure 1.1 Classification of web mining

applying standard database query mechanism and data mining applications to investigate it. The data on the web deals with unstructured data such as texts, semi-structured data such as HTML (Hyper Text Markup Language) documents, and a more structured data such as data in the tables or database generated in HTML pages, images, audios, videos, metadata and hyperlinks [KB00]. However, much of the web content is unstructured text data. Because of its immense size and data variety, a search query in the web IR results in thousands of web pages and most of them are irrelevant to the user’s needs. To prevail over this problem, web content mining provides a path to screen more specific data. The goal of web content mining from the IR view is to improve the information finding or filtering based on user profiles, from the database view is to model the data on the web and to integrate them so that more sophisticated queries other than the keywords based search could be performed.

Web Structure Mining: It is concerned with discovering the structural summary of the web page and the hyperlinks. Web content mining focuses on intra document structure that is within the document whereas, web structure mining tries to focus on inter-document structure that is within the web [CDG99]. In fact, there are several objects related in certain ways in the web. Hence, a peripheral and superficial application of the conventional procedures with an assumption of the existing independence of the web events may lead to disastrous and erroneous conclusions. This model can be used to categorize the web pages and make it possible to compare or integrate different web pages. It is useful to generate information such as the similarity and relationship between different web sites.
**Web Usage Mining:** It discovers the navigation patterns of the surfers from the web data. It deals with the prediction of the surfer’s behavior and interaction with the web. The web content and structure mining uses the primary data on the web, but web usage mining mines the secondary data, that is, the data from the web server, access logs, proxy server logs, browser logs, user profiles, user queries, registration data, user sessions or transactions etc. [CMS97]. The success of usage mining depends on what and how much valid and reliable knowledge one can discover from the huge raw log data [SCD00]. Web usage mining focuses on techniques that could predict user’s behavior while the user interacts with the web.

Web usage mining provides the support for web site design, personalisation server, other business making decision etc. Web mining applies the data mining to the web data and traces user’s visiting characteristics, and then extracts the user’s using pattern [QXW08]. Web usage mining process could be classified into two commonly used approaches [BL99]. The first approach maps the usage data of the web server into relational tables before an adapted data mining technique is performed. The second approach uses the log data directly by utilizing special pre-processing techniques. The applications of web usage mining could be classified into two main categories: learning a user profile or user modelling in adaptive interfaces and learning user navigation patterns [KB00]. Then the learned knowledge could be used for applications such as personalisation, system improvement, site modification, business intelligence etc.

### 1.5 PROCESS OF WEB USAGE MINING

Web usage mining process can contain three inter-dependent stages: data collection and pre-processing, pattern discovery and pattern analysis. In the *pre-processing stage*, the click stream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. In the *pattern discovery stage*, statistical, database, and machine learning operations are performed to obtain hidden patterns reflecting the typical behavior of users as well as summary statistics on web resources, sessions, and users. In the final stage of the process that is *pattern analysis*, the discovered patterns and statistics are further processed, filtered, possibly resulting in aggregate user models that can be used as input to applications such as recommendation engines, visualization tools, web analytics and report generation tools. The overall process is depicted in Figure 1.2.
The strategical execution of web usage mining is stated subsequently:

**Data Collection**: Data collection is the preliminary step of web usage mining. The data authenticity and integrity directly have an effect on the consecutive works smoothly carried on and the final recommendation of characteristic service’s quality. Moreover, the web usage mining technology observes the main data origin as the data from the web server, access logs, proxy server logs, browser logs, user profiles, user queries, registration data, user sessions or transactions etc.

**Data Pre-processing**: A number of databases are inadequate, inconsistent and incompatible as it exhibits noise. The data pre-treatment is to carry on a unified transformation to those databases. As a result those databases will get integrated and remain consistent. Perhaps, this helps in establishing the database that is potential enough to mine. The data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

**Data Cleaning**: The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. For instance, the cleaning process in a search engine query log implies the consecutive steps.
Creation of user profile for each user who set themselves to access textual information. This leads to the automatic removal of requested page comprising the file extension gif, jpeg, bmp (image file extensions) and others.

The elimination of the corresponding log entries when the requested pages fail to get loaded promptly due to some lapses in the process.

Removal of the log entries that consists of automated programs similar to that of web robot and so on.

Exclusion of those methods other than GET and POST existing in the log entries.

**User and Session Identification**: The task of this section explores and juxtaposes the users and their corresponding sessions from the original access log. User identification is identifying who access the web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user into individual sessions. A session is a series of web pages that the user browse in a single access based on a widely-used rule [WBC07]. The queries are split into two sessions if the time interval between them exceeds 30 minutes. Figure 1.3 explicitly brings out this process in detail.

![Diagram of User and Session Identification](image)

**Figure 1.3 Users and Session identification**

The rules adopted to distinguish the users and sessions are stated subsequently:

- Distinguishing every single web user through the usage of discrepant Internet Protocol (IP) addresses.
- Inferring the differentiation among the users through different browsers and operating systems in case of the utility of the same IP addresses.
- Deciphering the referrer information when the IP address, browsers and operating systems are observed to be the same. Further, the Uniform Resource Locator (URL) field
is checked, and a new user session is identified if the URL in the referrer URL field hasn’t been accessed previously, or there is a large interval between the accessing time of this record and the previous one if the referrer field is empty;

- The session may contain more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions.

Table 1.1 depicts the log information of a user and the user’s session information is displayed in Table 1.2 and Table 1.3.

<table>
<thead>
<tr>
<th>IP address</th>
<th>Date and Time</th>
<th>URL</th>
<th>Referrer</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:04:41</td>
<td>A.html</td>
<td>-</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:05:34</td>
<td>B.html</td>
<td>A.html</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:04:05:39</td>
<td>L.html</td>
<td>-</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:06:02</td>
<td>F.html</td>
<td>B.html</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:10:02</td>
<td>O.html</td>
<td>F.html</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:10:02</td>
<td>G.html</td>
<td>B.html</td>
</tr>
</tbody>
</table>

**Table 1.2 Session 1 of user 1**

<table>
<thead>
<tr>
<th>IP address</th>
<th>Date and time</th>
<th>URL</th>
<th>Referrer</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:04:41</td>
<td>A.html</td>
<td>-</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:05:34</td>
<td>B.html</td>
<td>A.html</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:06:02</td>
<td>F.html</td>
<td>B.html</td>
</tr>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:03:10:02</td>
<td>G.html</td>
<td>B.html</td>
</tr>
</tbody>
</table>

**Table 1.3 Session 2 of user 1**

<table>
<thead>
<tr>
<th>IP address</th>
<th>Date and time</th>
<th>URL</th>
<th>Referrer</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.168.1.41</td>
<td>25/apr/2011:04:05:39</td>
<td>L.html</td>
<td>-</td>
</tr>
</tbody>
</table>

**Knowledge Discovery**: Statistical method is used to analyse and mine the pre-treated data. Based on the user community’s interests an interest model is constructed. The typically used machine learning methods mainly have clustering, association, classifying the relation discovery and the order model discovery.

**Pattern analysis**: Challenges of pattern analysis is to filter monotonous information and to visualize and interpret the interesting patterns to the user.

**1.6 WEB SEARCH**

The web is an enormous, divergent, and vibrant storehouse of unimaginable treasury of information in the form of web pages. Due to the steady progress and proliferation of e-commerce, web services and web-based information systems, the volumes of click stream and user data collected by web-based organizations in their daily operations has attained culmination in extreme proportions. Figure 1.4 outlines the growth of active web sites from the year 1995 to 2013. Every month millions of web sites are newly added into the world of
information. Hence a proper tool is needed to search the information on the web. According to the survey done by Netcraft, Internet Services Company, England there is 739,032,236 sites in September 2013 which is 22.2 million more than August 2013.

The plentiful unstructured or semi-structured information on the web leads to a great challenge for the users, who hunt for prompt information. The scenario grows pathetic and traumatic especially for the business people, who need to provide personalised service to the individual consumers but buried in the billions of web pages. To overcome these problems, data mining techniques must be applied. The main objective of web mining is to develop more intelligent tools for the potential help of the user in finding, extracting, filtering and evaluating valuable information and resources from the web.

- **Finding relevant information**: People either browse or use the search service to find specific information on web. Today’s search tools have two problems. (i) Low precision due to the irrelevance of various search results. (ii) Low recall due to the inability to index all the information available on the web as some of the relevant pages is not properly indexed. This is a query triggered process.

- **Extracting new knowledge from the web**: This is a data triggered process. As it is hard to get relevant information, it’s very hard to make sense out of it.
1.6.1 Search Engines in Information Retrieval

Search engine retrieves significant and essential information from the web, based on the query term given by the user. Search engine plays an important role for the information seekers to provide the requested data in terms of web snippets. A web snippet denotes the title, abstract, and URL of a web page returned by the search engines.

The large size of the web has desired the continuous development of a class of systems that help the users search for documents on the web. These systems, known as search engines, give the users a simple interface by means of which to enter a set of terms known as queries. There has been enormous progress on the web search by using search engine. Starting from the text-based ranking algorithms, now complex ranking algorithms that use hundreds of features and much functionality based on usage data, such as spelling correction, query completion, query recommendations etc. Due to these advances, search engines usually satisfy a user’s informational need on well-formulated queries.

The user feedback cycle could be described as follows: users formulate queries in order to express their needs; queries are submitted to search engines; search engines suggest relevant documents which are recommended to users as answer lists for every submitted query; users select documents associated to their queries. Relationships among the activities of the user feedback cycle are shown in Figure 1.5.

![Figure 1.5 User cycle with search engine](image)

The retrieved result may not be relevant all the time. At times irrelevant and redundant results are also retrieved by the search engine because of the short and ambiguous query keywords [WJH01]. The user scans the search result from top to the bottom according to Joachim’s [J02] and then decides whether the web snippet is either relevant or irrelevant. A study done
by C. Silverstein [SHM98] on AltaVista search engine’s query log has shown that more than 85% of the queries contain less than three terms and the average length of the query is 2.35 terms. So the shorter length query does not provide any meaningful and relevant information to the users. [S08] reported that up to 23.6% of web search queries are ambiguous, this causes poor retrieval results.

Web is getting expanded day by day. Many search engines are used to retrieve the information from the web. In such situations, it is the responsibility of the service provider to provide suitable, significant and quality information to the user against their query submitted to the search engine. Query recommendation is an IR technology to recommend identical or related queries for a particular query [KYL12]. When the user is not satisfied with the result given by the search engine for the initial input query, query recommendation technique is used to provide suggestions to the user to frame relevant and meaningful queries in future to retrieve the relevant results. The recommendations given by the search engine depends on the real intent of the user, and the user’s intent is analysed from the search histories.

For example, a user who submits the query term “java” to any search engine may be studied. In this search, the user reviews the results relevant exclusively for “java bike” and not for the “java language”. Here, it is noteworthy that the user’s interest is on two wheeler vehicles particularly on “java bike”. At this juncture, query recommendation system best suits the need, as it could provide adequate suggestions about the bike if the same query “java” is triggered by the user anymore time in the near future. Figure 1.6 shows query recommendations provided by Google search engine for the query “inheritance”.

Web search engine stores the user’s navigational information in the form of search logs. Search logs are important repository to store the web user’s data. By mining the search logs, one can access the behavior and information needed by the user. Various algorithms and techniques are available to extract the useful knowledge from this kind of logs. Query log mining is one kind of web usage mining.

1.7 IMPEDIMENTS OBSERVED IN INFORMATION RETRIEVAL

The noteworthy drawbacks while the user uses the search engine for IR are

- Especially search engine users are inexperienced and they are usually casual users. They have very limited background knowledge about the domain they are searching for. Due
to the enormous growth of the web and its dynamic nature, short and ambiguous query keywords does not retrieve the relevant results properly.

Figure 1.6 Query recommendations

- Single user uses different computers for searching process and different user IDs are created for a single user. In this case, the query log processor creates a distinct entry with multiple IDs. Here the user’s real intent is not analyzed properly.
- If one system is used for searching process by different users then all the users have single user ID. In this case, multiple user behaviors are treated for single user.
- When a system connected with the internet, every time different IP is created and stored in the query log file.
- The identification of IP or user history is cannot be properly predicted as the user changes the system. In this case, the result may go wrong.
The traditional query recommendation techniques may go wrong for the above cases. The above anomalies are also considered and the solution is provided in this research.

1.8 SCOPE OF THE RESEARCH

At the end of the nineties, Bharat and Broder [BB98] estimated the size of the web to be around 200 million static pages. Gulli and Signorini [GS05] pointed out that the number of indexed documents in the web exceeds 11.5 billion. Now, the indexed web contains 2.14 billion pages (www.WorldWideWebSites.com). Exploring information through search engines is an important activity of the web users. User’s intent may vary from time to time and query to query. This research perceives that different queries from a user should be handled differently because a user’s preferences may vary across queries. For example, a user who prefers information about the fruit on the query “orange” may prefer the information about “Apple Computer” for the query “apple”. Interpret the human queries into search keyword is never straightforward [BCD09]. Especially search engine users are inexperienced and they are usually casual users. They have very limited background knowledge about the domain they are searching for. Due to the enormous growth of the web and its dynamic nature, short and ambiguous query keywords does not retrieve the relevant results properly. Query recommendation is a vital technique to provide the alternative queries as better alternatives instead of the initial query to get the meaningful and relevant information from the web.

The query recommendation technique suggests the user to explore information using appropriate and comprehensive queries to retrieve desirable and relevant results. Actually, the query recommendations of a search engine depend on the authentic and genuine intention of the user, significantly investigated from the search histories. For instance, a search pursued by feeding the query term “apple” obtains the exclusive recommendations for “apple iPod” and not for “apple fruit”. This recommendation is carried out with an assumption of the user’s interest fixed to “apple iPod”. At this juncture, with reference to the previous probe of the same user on “apple iPod”, query recommendation system attempts to supply the results if the query term “apple” is tried by the same user next time. Here, the recommendation gets specified with reference to the user’s previous explorations. Usually, a query log file comprises a log entry for every single query of the user in the search engine. An effective strategy of mining the query log file improvises the effectiveness of the search engine in order to make personalised recommendations on items.
1.9 PROBLEM IDENTIFICATION

Information searching and retrieval is an important activity of the web users from the web repository. But there are some difficulties faced by the web users. The major problems in the IR are simple and ambiguous query keywords, information overload and user’s real interest about the searching session is not cached properly. Hence the retrieved result is not relevant and meaningful at some times. So, the research for providing better alternative queries is an important one. The recommended keywords are relevant to initial query, accurately catch the user’s intent and produce the diversified results. Several algorithms for query recommendations exist in the history with some drawbacks. These problems are addressed in this thesis. The proposed work recommends the queries in web IR process as a personalised one. That is the recommendations are based on the real search intent of the user and increase the number of relevant web snippets appeared as the result of the searching process.

1.10 OBJECTIVES OF THE RESEARCH

The hypotheses framed for this research are enunciated subsequently:

- To pursue an empirical study on query log to observe the search behavior of the web users. This is recognized through the occurrence of frequent access patterns that correspond to the calculation on the count of clicks done on the URL.
- To enhance the query recommendations by combining the similarity between the query keywords, clicked URL and above all the concepts retrieved for each query. Both the positive and negative concepts preferred help to explore the string of similarity between the concepts generated which in turn leads to cluster the users with similar intentions further, leading to make recommendations.
- To formulate a time-confined query recommender system, the time variant and invariant query clusters are generated; this cluster contains the similar queries used to provide the recommendations. This technique is an amalgamation of content-based and collaborative approaches.
- To explore the favourite query of the user, to fix the user’s real search intention with the help of the hybrid user profile and to prioritize the recommended queries.

The crucial aspiration of this work is to formulate a new strategy to provide alternate queries for each individual input query. The tetra-faceted contributions of this research are
Phase I: Innovative Strategies of Query Log Analysis
- Embarks on Identifying the Users and Sessions from Search Logs
- Deals with the Frequent Web Access Pattern Mining for Web Users
- Fixes the Count of URL Clicks from the Search Engine Query Logs

Phase II: Time Invariant Query Recommendations from Search Engine Query Logs
- Describes Time Independent Query Recommendations using Query and URL Similarity
- Exhibits Time Independent Query Recommendations using Query, URL and Concept Similarity
- Expresses Time Independent Query Recommendations using Positive and Negative Concept Preferences with Query and URL Similarity

Phase III: Time Heuristics Query Recommendations and Re-ranking of Recommended Queries
- Describes the Time Heuristics Query Recommendations
- Deals Time Heuristics Ranking Approach for Recommended Queries

Phase IV: Personalized Query Recommendations using Concept based Hybrid User Profile
- Delineates Concepts Extraction
- Depicts Query Recommendations using User Profile and Ranking of Queries

A global view of the relationships between data mining techniques used in this thesis and the applications generated from the use of these techniques over the query log data is given in Figure 1.7.
1.11 ORGANIZATION OF THE THESIS

The prime part of this research is oriented towards analysing the behavior of the search user using search query logs. The thesis provides a set of queries as recommendations to frame the better query in future, based on the user’s intention in the IR process. This thesis entitled *Time-Confined Search Query Recommendations in Web Information Retrieval using Query Logs* consists of eight chapters. The proposed plan of this thesis is shown in Figure 1.8.

**Chapter 1 - Introduction** provides a precise outline on web mining and search query recommendation techniques. This chapter elucidates the issues in web IR process, in addition to the purpose of query recommendation system in the search process. Besides, the demand and application of recommender system in day to day search activities for IR and the scope of the research clubbed with its methodology are dealt in detail.

**Chapter 2 – Review of Literature** exhibits the existing search log’s analysis and the behavior of the user in the searching process. This chapter makes a survey of all apposite literature and reviews of discrepant techniques in query log mining and recommendations juxtaposed with its interpretations clubbed with the tools deployed. Invariably a holistic evaluation of different ranking techniques along with its applied parameters to prioritize and re-rank the recommended queries is elaborated here.

**Chapter 3 – Explorative Approach of the Proposed Research Methodology** elucidates the holistic design of this research, its scope, objectives, problem specification and methodology. This chapter serves a preamble for the typical research pursued in query logs, prioritizing and shortlisting the recommendations. Exploring the user’s intention and providing a time-confined effective strategy in the rest of the Chapters from 4 to 7.

**Chapter 4 – Innovative Strategies of Query Log Analysis** fixes its crux on the identification process of web information seekers juxtaposed with their respective search sessions and their navigations from the search log file. With the inference made through the users’ search history besides, their respective navigational aptitude the frequently accessed pattern for each query is generated. From these frequent patterns, the URL click count is typical which is calculated through trivaricated techniques namely, Sequential URL Click Count, Bit Vector URL Click Count and Hash based URL Click Count. These proposed techniques are compared and evaluated using the sample log file in the experimental results section.
Chapter 5 – *Time Invariant Query Recommendations from Search Engine Query Logs* edifies the deployment of query recommendation strategies in the process of web exploration, that is not confined by time. In fact, this time-independent strategy simplifies the
cumbersome search expeditions of the users. This chapter is packed with information on queries and its navigations occurred at discrepant time spans. The strategic planning of combining and calculating the similarity between the query terms and clicked URLs for the given query to provide a suitable recommendations for the initial input query is described here. This collaborative application of recommendations technique helps to observe the users with similar interest and to cluster them. The cluster provides the recommendation for the similar users. The expert’s query term is provided as the recommendation for naive users. Apart from the query keyword and clicked URL similarity, the concepts available in the clicked web snippets are also considered to endow with recommendations. This combined similarity provides the query recommendations. Moreover, this chapter consolidates the experimental results with an AOL search log to edify the time independent similarity.

**Chapter 6 – Time Heuristics Query Recommendations and Re-ranking of Recommended Queries** highlights the recommendations technique with a special perspective to potential recognition of the queries; arranging the successive occurrence of the queries into time-bound clusters that exhibit its fulcrum on the hub, authority and time measure; Pre-processing of query terms and identification of frequently accessed queries and URLs from the query log file by considering the prefix and suffix patterns of the queries and clicked URLs. The application of modified frequent pattern generation algorithms namely Prefix Span [PJB04] and Up Down Directed Acyclic Graph (UDDAG) [C10] are also discussed. Next hub and authority weights are calculated for the frequent items. Here, the clustering of time-bound similar queries is pursued through Hierarchical Clustering Algorithm besides, assigning the weight t-measure. With an assumption that the recent occurrence of similar queries boost-up a better recognition when compared to the earlier queries this concept is presented. The cluster generated in this approach is used to provide query and URL recommendations to the user.

Four different cases of search user $U$ and query $Q$ are identified, such as;

**Case 1**: The query log may comprise $U$ and $Q$ that is absolutely new to the search process.

**Case 2**: $U$ is new and the query keyword $Q$ may exist already.

**Case 3**: $U$ exists previously in the query log however, $Q$ is new.

**Case 4**: The existence of $U$ and $Q$ in the query log.

Ranking order of queries is compared based on the rank measures Euclidean Distance, Manhattan Distance, Area under Curve (AUC), Ordered Area under Curve (OAUC) and
Accuracy. The ranking order is also measured using Mean Reciprocal Rank (MRR) measure. Finally the proposed method has been evaluated using real data set of AOL search engine query log.

Chapter 7 – Personalized Query Recommendations using Concept based Hybrid User Profile overtly presents the pragmatic use of query recommendation technique with a targeted perspective to the query log, user log and concept log. The user profile seen in this chapter is a typical hybrid application with a conglomeration of both implicit and explicit feedback information. Implicit information is retrieved from the concepts of clicked web snippets and the user’s navigational behaviors. Explicit information is gathered through feedback forms. The important concepts based on their support value is identified from the clicked web snippets and it is stored in the concept log. These log files are used in the query recommendation process. The recommended queries are re-ranked by showing the user’s interest in the clicked snippets. The experimental results with real data set shows the recommendation process of the proposed method. Finally the recommended queries are evaluated by assigning the relevancy scores 0, 1 and 2 by the users. The similar users based on the relevancy score is clustered. Kappa measure and coefficient of variance are used to identify the similarities between the users.

Chapter 8 – Conclusion is edified through the proposed approach the possibility of a time confined query recommendation in web IR through the collaboration of concept based hybrid user profile. This study obviously highlights the avenues of prospective developments in recommendation systems to improvise the efficacy of every search. Time variant is given due importance significantly with an aspiration to provide a novel query recommendation strategy known for its limitless efficacy. Precisely, the futuristic scope of this concept based hybrid user profile with time confined widens several territories of research requisites besides reinstating the demand for quality research in field of query recommendations.

1.12 SUMMARY

This chapter summarizes the introduction about web mining, categories of web mining and the process of web usage mining. Different recommendation techniques and search process in the IR, search personalisation and issues in the query recommendations are identified and discussed. Scope of the research, problem identification, objectives of the research and different phases in the proposed research methodology and organization of the thesis are described.