A CONTEXT SENSITIVE AND PERSONALIZED QUERY AUTOCOMPLETION TECHNIQUE

MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF

Master of Technology

In

Information Systems

Submitted By:

ANTRA KATIYAR

(2K15/ISY/06)

Under the Guidance

Of

Dr. O.P. Verma

(Professor, Department of Computer Science and Engineering, DTU)

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY

(2015-2017)

CERTIFICATE

This is to certify that **Antra Katiyar (2K15/ISY/06)** has carried out the major project titled "**A Context Sensitive And Personalized Query Autocompletion Technique**" in partial fulfilment of the requirements for the award of Master of Technology Degree in Information System during session 2015-2017 at **Delhi Technological University**.

The major project is bonafide piece of work carried out and completed under my supervision and guidance during the academic session 2015-2017. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/Institute for the award of any degree or diploma.

Dr. O.P. Verma

Professor

Department of Computer Science and Engineering

Delhi Technological University

Delhi-110042

ACKNOWLEDGEMENT

I take the opportunity to express my sincere gratitude to my project mentor Dr. O.P. Verma, Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi, for providing valuable guidance and constant encouragement throughout the project. It is my pleasure to record my sincere thanks to him for his constructive criticism and insight without which the project would not have shaped as it has.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.

Antra Katiyar

Roll No. 2K15/ISY/06

M.Tech (Information System)

Department of Computer Science and Engineering

E-mail: antrakatiyardtu@gmail.com

ABSTRACT

Query Autocompletion is a leading attribute of Search Engines which makes the user's search experience better by predicting the query. QAC methods suggest query suggestions to users, after they enter some of the keystrokes in the search engine. This is done by predicting the query using past query logs and other trends. Current QAC methods use the Most Popular Completions as the suggestion results. Context and Personalized techniques are proposed already but they are used separately. The present methods being incorporated are the location and past searches sensitive QAC.

In this proposed work of thesis, we will talk about a hybrid technique by combining both the context sensitive, trending and personalized suggestions. The improvements which are made in the base paper are that a new approach can be proposed by combining the three techniques to create a hybrid technique. It intends to incorporate three major research works: **Time sensitive** (based on time series and trends), **Context Sensitive** (based on recent searches done) and **Personalized** (based on gender, location and age-group) query auto completion. Thus an algorithm that considers all these parameters will be better at predicting the user query. The results predicted are better in reducing the user keystrokes during the search and also reduces the searching time, and also enhances the reliability of the search engine.

Further improvements can be done by extracting the user's browsing history to determine keywords, interests and other user-specific data for enhancing the result predictions.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABBREVIATIONS

QAC	Query Auto Completion
MPC	Most Popular Completion
AOL	America Online Query Log
CS	Context Sensitive Technique
PS	Personalized Score
HC	Hybrid Completion
TS	Time Sensitive

Chapter 1

INTRODUCTION

Search engines receive millions of queries daily which are interrelated depending on the user. The queries are recorded and saved for future references and contribute to a huge database. This database can be analysed or mined to find out the trends in user queries and how they vary with time, location and other parameters. This is already being done by Google in the form of Google Trends. It offers deep insights into how and what people from where are searching.

The proper analysis of the user searches can present us with better results for the suggestions in the web search engine when a user is typing just some of the initial keywords of the related search. A very straightforward approach is Most Popular Completion. The most popular searches are presented in the drop down list with the most popular one on the top of the list. It means the searches which are most popular and done by the people related to the entered keystrokes in the search engine.

Though this method is simple and easy to implement, it is quite inefficient and time taking for the less common queries. Thus other methods and criteria need to be incorporated into our query auto completion technique in order to improve the accuracy of our prediction.

Modern search engines such as Google and Bing take into account the user location while returning the suggestions. Thus the predictions vary according to the country of the user. Further research into the field has found that time specific trends can be used as a criteria to improve suggestions. Similarly, users past searches may be helpful in cases the query being predicted in resulted to the context.

Even better predictions can be produced if we have the details for user's age, gender, interests and may be with consent, browsing history and keywords.

1.1 What is a Query?

A web search query is the keywords entered by the user on the web search engine for the search process. According to the user needs, some keywords are entered to view the results.

1.1.1 Types of Queries

A web search query can be categorized into three types of queries:

- **Informational Query (Do)**: Queries which gives a thousands of results as it covers a wide range of data. (e.g cars, trucks, bikes).
- **Navigational queries (Know)**: Queries that demands to a single entity or a website of a particular entity (e.g google drive , google maps, SBI online)
- **Transactional queries (Go)**: Queries that shows an action of the user related to a particular query like booking a ticket for airline or downloading an IEEE paper or buying a smart phone.

1.1.2 How Does It Work?

Autocomplete or Word Completion is a technique in which the interested user inputs or types the initial few letters of the word search. The words with the highest probability of searching with the given letters in the search box are shown in the drop down list. The user can select the given specific word with the up and down keys. If the given drop down list does not match with the intended query user wants to search the user enters some more letters for the better suggestion list. With every keyword entered the drop down list modifies itself so that the user selects from the given suggestions and saving the time and fulfilling the use of a search bar rather than writing out the whole query in the search bar. Autocompletion uses the language modelling technique in which with the given set of letters the most popular words which are searched are previously calculated. Recency model is also used in combination with the language modelling so that the queries which are used more recently are also predicted. User can also enter the words into the prediction dictionary.

1.1.3 Examples

a) Google Instant

Google Instant was introduced to increase the effectiveness of the search engine. It helps in predicting what the user is searching for and shows as the suggestion list. It uses the Google Autocomplete technology for the prediction of the user query. It shows the top 10 completions with the initials as entered in the search bar.

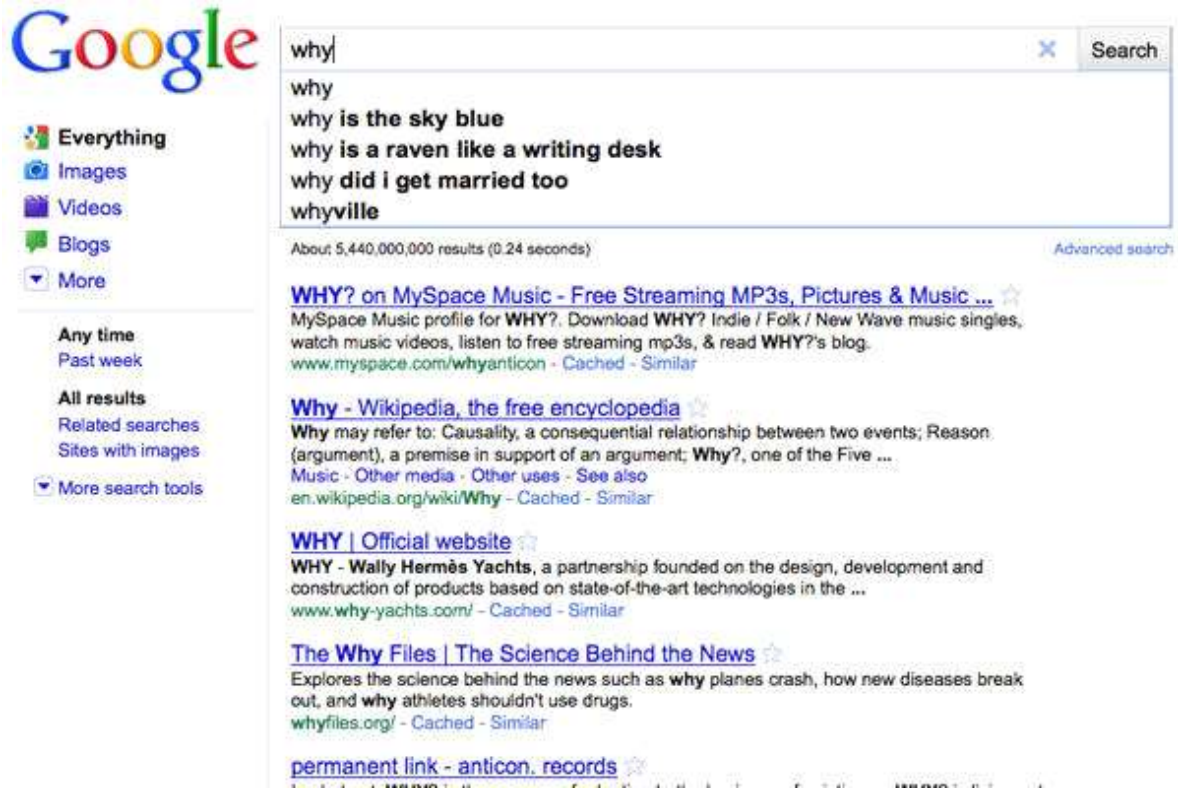


Fig 1. Google Instant on typing “why” as the initials

The above example of Google Instant shows the top 10 suggestions when the user enters “why” as the prefix of the searching query. The suggestions are shown and the user selects the one of the result of his/her interest. It helps in getting to the right content much faster. The user can also enable or disable the Google Instant from their search engine depending on their use

b) Bing Autosuggest

Autosuggest is the autocomplete technology used by the Bing. Navigation and Search history patterns are used to calculate the "Degree of Confidence" value on which the results appear in the drop down list. A blue box also appears with the text of the most popular query so as to speed up the time to find the required result.

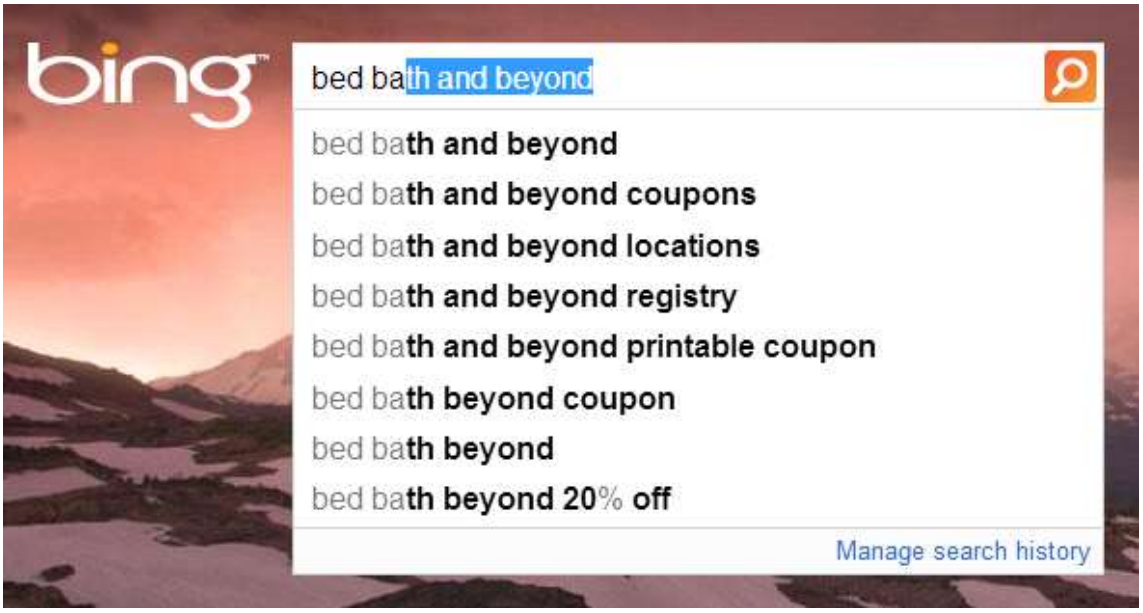


Fig 2. Bing Autosuggest on entering "bed ba"

The above example of the Bing Autosuggest shows the possible results when the initials of the query "bed ba" is typed as initials in the Bing search engine. The blue box shown with the most possible return query as text in it is called as Ghosting Technique. The result with the highest confidence value is ghosted in the search box . It helps in formulating the user query result faster before entering the search key.

1.1.4 Advantages of QAC

- **Faster Searches**: It helps in predicting the search before the user enters complete query and saves at least 2-5 seconds per search.

- **Smarter Predictions**: When user don't know what exactly he is searching for the auto completion helps in predicting all the possible matches for the given initials and user can stop typing the keywords.
- **Instant Results**: User starts typing the query prefix and as soon as keywords are entered the results are shown instantly in the drop down list rather than the user typing the whole query and then the accurate results are predicted.
- **Reduced Keystrokes**: When user enters the initial one or two characters the results are predicted which helps user in typing the less number of characters and saving the time.
- **Query prediction and correction**: User may not know the exact searching query so with the results he can refine the queries to get better predictions.
- **Better user experience**: A good QAC helps in suggesting accurate results before the user types the complete query and fulfils the auto completion technique of the search engine.

1.2 OBJECTIVE

To enhance the performance of the search engines so as to improve the predictions of the suggestions which are shown in drop down list to increase the efficiency of the search engines and reduce the keystrokes related to the search.

1.3 MOTIVATION

In a web search, when a user enters the keywords in the search engine the results are shown in the drop down list. The drop down list show the suggestions depending on the mostly searched queries to the less searched queries. The most common technique which is used for the suggestions relate to the Most Popular searches done by the people in the past. The past searched queries provides a database for the current searches. Some other techniques were also introduced like time series and the personalized suggestions which were not combined.

This thesis includes a hybrid approach of these three techniques of the Most Popular Completion, Time sensitive and the Personalized suggestions. Most Popular Completion

technique uses the previously searched queries which are mostly searched according to the given keywords. Time sensitive technique uses the current trends. The trends are fetched and the results are combined with the Most Popular searches. The Personalized suggestions helps in providing the suggestions as user specific. Previously, the techniques which were used mostly provides a global result which were not user specific. This hybrid approach stipulates a user specific suggestions depending on his/her interests which are combined with the current trending topics and the most popular searches. The combined results provide better suggestion list. A personalized QAC list is shown which combines the three techniques. This hybrid approach also improves the QAC effectiveness and enhances the working of the web search engine.

1.4 GOAL

This thesis work introduces an improved methodology for the suggestions or we can say the possible completions of the initial prefix entered in the search engine which are shown as a drop down list in the web search engine query. Previously, the most popular searches (wisdom of the crowd) are preferred to be shown in the results with mostly searched query as the top suggestion of the list. Time sensitive and personalized query auto completion techniques are also suggested. In this work, we will talk about a hybrid approach by combining both the time sensitive, context sensitive and personalized suggestions. The improvements which are made in the base paper are that a new approach can be proposed by combining the three techniques to create a hybrid technique. It intends to incorporate three major research works: **Time sensitive** (based on time series and trends), **Context Sensitive** (based on recent searches done in the previous hour or half an hour which can be related to the current query) and **Personalized** (based on gender, location and age-group or some user likes taken from some Facebook liked pages) query auto completion. Thus an algorithm that considers all these parameters will be better at predicting the user query. The results predicted are better in reducing the user keystrokes during the search and also reduces the searching time, and also enhances the reliability of the search engine.

1.5 ORGANISATION OF THE THESIS

Chapter 2 includes literature review of Query Autocompletion Techniques. The formulation of the query suggestion techniques using geographical locations, time series analysis, seasonal queries, detecting the current trending topics and personalized results are shown. All the techniques uses different factors and scores to find the most specific results and ranked according to different criteria from the most appropriate one to the least appropriate. It shows how the modifications over the years have been done to increase the quality of the effectiveness of the Autocompletion techniques.

Chapter 3 includes the models and methods which are used in the thesis work for the new Hybrid technique. We will start with the basic model of the Most Popular completion technique which is then combined with the Context Sensitive as the first step for the proposed model. Then as we move along to develop the proposed model the Personalized technique has been incorporated to provide the user based results.

In Chapter 4, the proposed work is discussed and the implementation of the proposed model is done on the AOL data set.

Chapter 5 analyses the proposed work with the experimental results which is obtained by the implementation. The results are efficient, user specific and more reliable as compared to the previous results.

In Chapter 6 the Conclusion of the proposed model in the thesis work have been discussed with the future work as what further modifications can be done for the improvisation of the current results and which could prove to be the important consideration factors for the predictions.

Chapter 2

BACKGROUND WORK

In this chapter, we present all the findings from the internet and other sources that shed a light on the past, present and future query auto completion along with the researches being undertaken to make it more accurate.

Before this, we intend to find the usefulness and need for research and development of Query Auto completion and how it is fruitful. Why do sites intend to use and improve it and why it matters to them?

We'll also explore the challenges we face while developing a query auto completion technique and the obstacles we need to overcome.

How important is query auto completion to the user and how it helps the user in writing the query and the predictions of search engine.

2.1 CHALLENGES

- **Huge Database:** Query databases are considerably huge and deciding the suggestions from such a large number of possibilities is a major challenge in itself.
- **Time Constraint**: Unlike search results, query suggestions need to be sent back immediately with minimum time lag. Thus the algorithm should be fast enough to cope up with the user's typing speed.
- **Relevancy**: Most queries having high frequency in the database are irrelevant to the correct context and should be smartly ignored by the algorithm.
- **Safe Search**: One more thing to consider is not to suggest offensive or inappropriate words which may be quite common in the database.

2.2 RELATED RESEARCHES

Considerable research has been done to suggest better queries, below is a summary of researches conducted by various scientists.

Auto-completion is a well-known element and has been used in a few applications going from early UNIX Shells to present day content tools and web programs. Past work on it can be assembled into two primary classifications; the first gathering which sends data recovery and NLP systems to produce and rank suggestions when the user enters new words and characters. For example, sentence completion is done in view of vocabulary insights of content accumulations and positioned AC hopefuls as indicated by a recommendation model trained by Latent Dirichlet Allocation (LDA) [12]. It built up an ongoing inquiry development framework that creates a new dropdown list of possible solutions. In the other group of AC techniques which included the above work queries are computed and their values are stored in a hash table for the lookup search. The list of queries are modified according to the new entries or query searches done by the user.

The users are filtered according to their prefix matching condition. The positioning in query systems depends on static scores that are derived on the basis of their significance value. For instance, in an item motor, for example, ebay.com with items names as user name, a static value can be allotted according to the indications by ubiquity, cost or the survey scores of items. With regards to web seek, the most regular approach is to rank hopefuls (question proposals) as per their past fame.

Most Popular Completion (MPC) is considered as the most basic technique and it is considered as the greatest probability estimator for a web search [3]. The authors also enhanced a setting mindful system in which the assigned scores value for the users are joined with relevant scores in light of late according to the user history to process the final positioning. This thesis work is based on all the previously mentioned techniques. Our work can be likewise joined with the hybrid structure to enhance the positioning of competitors with no or little setting [3].

Web search also considers the geographical locations for query formulation. It helps in finding the better predictions for the informational and navigational queries. The geographical

queries helps in predicting location specific advertising. The user locations and previous location specific queries are used to return the in the most relevant suggestions [16].

The data needs and web queries are changing continuously according to the world needs which are reflected in the web search. The modifications and the new additions in web searches have been investigated and stretched out the displaying structure to fuse the time removed from record time-stamps. For the correct prediction of the query in the prediction list is calculated according to the time stamp values [11]. Their theory was extended which utilized the substance of best-positioned records for approximating the most related time-intervals for the queries. The changes in the data store values and its consideration in the query suggestions are related to each other. A model considering the values was proposed which defines the changes according to the change in query inputs.

Relatedly, another model researched utilizing straight time-arrangement models for term weighting which presented a time-sensitive model of PageRank which can be used as an efficient score value for archive positioning [9]. Another model coordinated in figuring out how to rank by presenting better approaches for upgrading for both freshness and significance which classified queries into different classifications in view of their change of prominence after some time [13]. The authors demonstrated that checking the query updates and its popularity can uncover helpful signs for recognizing the adjustment in query expectation. They likewise classified questions with the understood worldly goal (e.g. Halloween) as indicated by their past events in the logs.

Shokouhi utilized time-series analysis as a strategy for ordering seasonal query and showed that queries with comparable seasonal patterns which are related to the lexical values [4] [9]. Similarly it presented a unified display for query prediction where the query count for each query gets affected by the frequency for the similar query inputs.

Another model built up using a compact model for time-arrangement and recommended a technique for distinguishing considerable changes in the query count [8]. It researched the effect of changes in query flow and the evolving patterns in the query. They contended that creating the diagrams without any preparation is not possible. So they evolved a new incremental technique for effective suggestion list. They grouped the queries as indicated by their recurrence time-

arrangement. They proposed that their resemblance factor can be utilized for the classification purpose and the plausible results.

Another strategy reranked query on the basis of a time stamp associated with them in bits and proposed that their resemblance can enhance efficiency for queries [11]. Radinsky et al. have played out an examination on consistency of distinguished users practices, (for example, inquiry clicks) utilizing time-arrangement analysis and a learning model. They additionally think about the general relevance of time arrangement analysis for displaying query patterns. Be that as it may, those investigations have been performed on minor arrangements of the query (≈10,000 questions) and of brief timeframes (6 months), prohibiting identification of a query forecast in a long period. In their work, they examined a huge scale data of around 1000k questions over a time period of approximate 5 years with the end goal of efficient query results.

Another interesting approach is the trend detection in query. It includes the periodic behaviour of the searching query [7]. The trends are placed on a two dimensional axis where the ends of the axis defines the highest query and the query which are rarely searched. The centre of the axis defines the queries that are popular at times depending on the time. Like some of the activities are leisure time activities and these queries are searched periodically i.e. on weekends or public holidays like the fun parks, sale in shops etc. The other activities are the trending topics at time which have gained a sudden peak in their search. For example as the cricket scores websites at time of World Cup or news regarding any attack in any part of the world which suddenly attracts human attention.

To make the query results more efficient and personalized according to the previous searches of the user. It means using the user past queries for estimating the current predictions so that the results are not the generalized one but they are modified and altered according to every user query request. This introduced a new concept of Personalized QAC.

Personalized QAC predicts the result on the basis of the likelihood of the user past queries [14]. A tree data structure is used for the storage of past queries which are ordered according to their past frequencies. It is an efficient look up technique for the user interested results. Two types of user history long and short term are used. Long term are beneficiary at the starting stage while short term are beneficiary for the continuous searches during a session. The

entire past search of the data is considered as a valuable database for the initial query purpose and the searches done in the previous context that is either an hour or an half hour basis that continues to improve the corresponding next query in that session.

A Query expansion technique is also considered effective during the query formulation phase by the user which is known as Real Time Query Expansion (RTQE) [6]. It considered the three phases for the query expansion. Baseline interface where no expansion for the query is provided then the Real Time phase where user is entering the query and the expansion is provided in the search bar and the last phase Retrospective where, when the user has written the complete query in the search box and enter the search button for the query suggestions.

For a known query the Baseline is faster than both where Real Time is faster than the Retrospective. For exploratory query the real time is faster than retrospective which is faster than Baseline.

The queries based on trends are most likely related to the searching query. The Time Sensitive QAC depends on the time when the search is done [1]. Recency and the seasonality are the two factors which help in determining the result prediction. Time series analysis turns to be a better factor for determining the queries which are searched during the weekends, queries which are trending during the summer vacations or the holidays. Time series considers the recent variations as well as the long term observations. The scores of both the factors are considered for predicting the queries on the next day.

Time sensitive QAC uses the database which is previously computed on the basis of the popularity factor and when these prefixes are entered in the search box they are dynamically altered depending on each user i.e. personalizing the query results [5] [1] [8]. Google trends shows that how query is majorly dependent on the time factor of when the query is being done. Google trends results are also modified during these timings like New Year's Eve or Christmas.

The queries are returned according to the Google trends and the cyclicity of the query which then considers the factor of the user previous queries [5]. The context of the user is incorporated into the results for the further modified results in the drop down list.

The concept sequences mapping is also done for the improvement of the results. It takes into two steps for the context factor [15]. The two factors are the offline and the online mode of sequences. The offline mode maps the concepts into the clusters form and when the user enters the query the online mode maps the offline concepts dynamically with the user queries in the past session which acts as a good context and improves the usability of search engines.

Chapter 3

MATERIAL AND METHODS

In this chapter, we will discuss additional capabilities of QAC techniques and we will also review possible approaches and their advantages for using in QAC model.

3.1 TIME SENSITIVE QAC

QAC is an important component in the web search engines. It raises the look and understanding by helping the clients in giving the suggestions before writing the whole query in the search bar. The QAC techniques which are used nowadays, rank recommendations as indicated by their past prominence. Nonetheless, inquiry prominence changes after some time, and the positioning of required query must be balanced as according to the new changes. For example, while Diwali may be the correct proposal in the suggestions list of writing di in October, dictionary may be better on the other time. With the approach of QAC as the main factor in the consideration for the web search engines and the efficient results its rising importance is considered.

One step further another approach is considered as Time Sensitive QAC. Rather than positioning hopefuls as per their past prominence, time series analysis and frequencies are the major factors in contribution for the result. Our investigations on approximate 1000k questions and their day by day frequencies examined over a time of 5 years demonstrate that foreseeing the prominence query exclusively on the basis of their previous evaluation is deceiving, and the figures acquired by time-series analysis are more relevant. The outcomes additionally propose that displaying the patterns of queries globally can fundamentally enhance the positioning of QAC strategy.

3.1.1 NEED FOR TIME SENSITIVE QAC

QAC is a technique which helps in determining the user's intention of the query he is searching for in the search engine suggestions list before the user writes the complete query. As the user

starts with the initial key values of the query the matching list according to the prefix are evaluated. The queries are suggested according to the Most Popular queries but the time factor is also considered [1] [4]. The Past frequencies plays a major role for determining the future results. But this approach is static in nature as the data needs are changing time to time so these modifications are also need to be considered which helps in providing more specific and appropriate results.

Fig. 3 Google predictions for the prefix 'di'

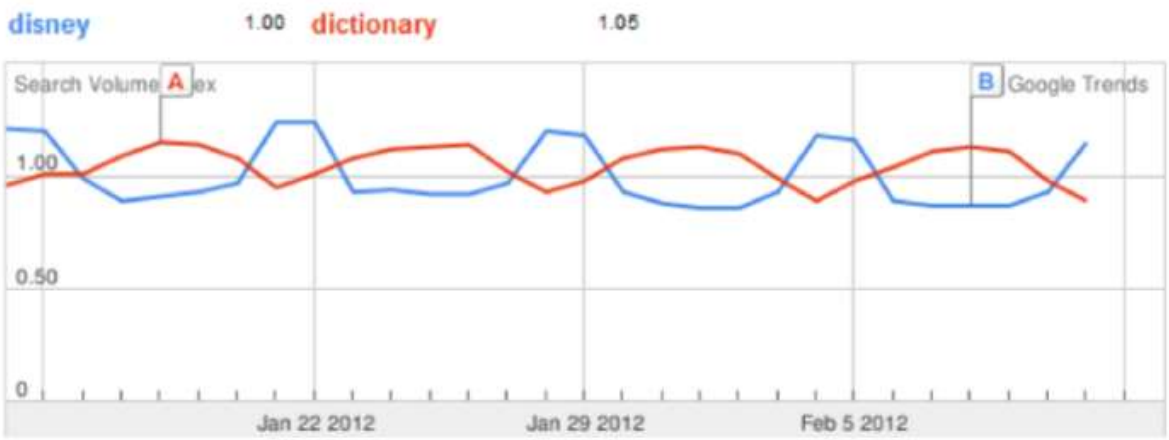


Fig 4. Daily frequencies for queries dictionary (red) and Disney (blue)

For Example, Consider in figure 1 where a user has typed "di" so we can see that the results are shown where the Disney and dictionary are the results. But the major factor which needs to be considered is the time when the query is searched [4]. On weekdays dictionary is more appropriate suggestion whereas on weekends Disney has a higher score than the dictionary. Similarly if the query is searched in the month of October Diwali is the most appropriate result. So the different time spans are need to be evaluated and their contribution must be considered in the results.

The first problem is that if we ignore the time related variations then the query results are much better if we consider the shorter time spans as compared to a larger history. Then the time related analysis is also combined with the popularity results and the results are dynamically evolving. This technique is performed over a large database on a weekly and monthly interval which then showed a high quality ranking results.

3.2 CONTEXT SENSITIVE QAC

It is shown that when the prefix or the input typed by the user is just a few letters then the suggestion list is quite not reliable. So the Context Sensitive technique can help in providing better results in the drop down list. In this technique the user's context is an important factor for future forecasting. A new technique proposing Nearest Completion, which yields the results of the user query that are most likely. The AOL log was used to perform the experiment and the results were evaluated. It was shown that when the query done in the previous half hour or one hour time interval is anyhow related to the present query means that the context is relevant to the current query then the results are more appropriate. But when the Context is not associated with the current searched query then the score for the completion is zero. To solve this issue, a Hybrid approach appeared to command both the techniques of Nearest Completion and the Most Popular Completion which accomplished a change in the score as compared to the Most Popular technique.

3.2.1 HOW IT WORKS

The Context Sensitive approach uses the Maximum Likelihood property. It is a conditional probability estimation for the current query. It tries to predict the future query according to the context of the user. It estimates that with the given context what is the probability that the next query will be q. The results are mapped with the dataset in the documents and their weightage value is calculated based on which the query forecast is done. For the prefix storage a tree data structure is used. The prefixes are stored on each node and then used for an effective lookup which is same for all the users. It is precomputed and then dynamic modifications are done according to the new changes.

3.2.2 COMPARISON

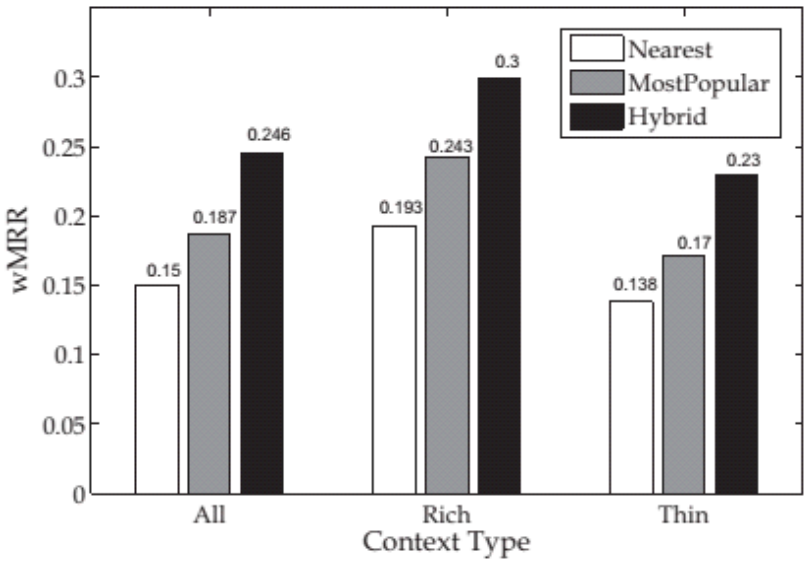


Fig 5. Weighted MRR of the 3 algorithms

Consider Fig 5 where the results for the three techniques are compared. When the context is a mixed combination of relevant and irrelevant queries. Most Popular Completion is better than the Nearest Completion in all the data sets [3]. But the combination of Most Popular and Nearest

Completion which is Hybrid is most effective in all the cases. When the context is rich in nature the Hybrid tends to provide the most effective and relevant results.

Context	Query	MostPopularCompletion	NearestCompletion	HybridCompletion
french flag	italian flag	internet, im help, irs, ikea, internet explorer	**italian flag**, itunes and french, ireland, italy, irealand	internet, **italian flag**, itunes and french, im help, irs
neptune	uranus	ups, usps, united airlines, usbank, used cars	**uranus**, uranas, university, university of chicago, ultrasound	**uranus**, uranas, ups, united airlines, usps
improving acer laptop battery	bank of america	**bank of america** , bankofamerica, best buy, bed bath and beyond, billing	battery powered ride ons, battery plus charlotte nc, battery died while driving, best buy, battery replacement for palm tungsten c	**bank of america** , best buy, battery powered ride ons, bankofamerica, battery died while driving

Table I. The top 5 completions provided by the 3 algorithms

In Fig 6 when the query prefix is 1 or 2 letters then the Nearest completion provides the most appropriate suggestion list and the Most Popular provides the results according to the popularity of the given prefix. Like in the initial two queries when the context is useful for the current query the MRR of the Nearest and Hybrid is much better than the Most Popular [3]. When the query context is irrelevant then the Most Popular is the best fit technique. But for both the cases Hybrid provides the suggestion list in an effective manner since it combines the list of the above two techniques.

3.3 PERSONALIZED QAC

This technique was proposed for personalizing the query predictions which suggests the results according to the specific user. It is like a training model where the user data is used as a database for the learning process which then helps in better predictions. Some query searches are specific to the user which can't be predicted with the Nearest or Most Popular Completion results [14]. This predictions are done by personalizing the suggestions may be according to the male or female query or the user likes and dislikes.

Initial, an arrangement of special impressions are tested from the logs. Every impression comprises of a one of a kind client ID, a period stamp, the submitted question, and the arrangement of results displayed to the client alongside data about understood measures, for example, the user ticks and their stay time on a query. Further, records that got a sufficient ticks are marked with important codes and others are viewed as irrelevant. The SAT-click is theoretical concept, but it is assumed that if the stay time on the last query is more than 30 then it is considered as SAT.

3.3.1 NEED OF PERSONALIZED QAC

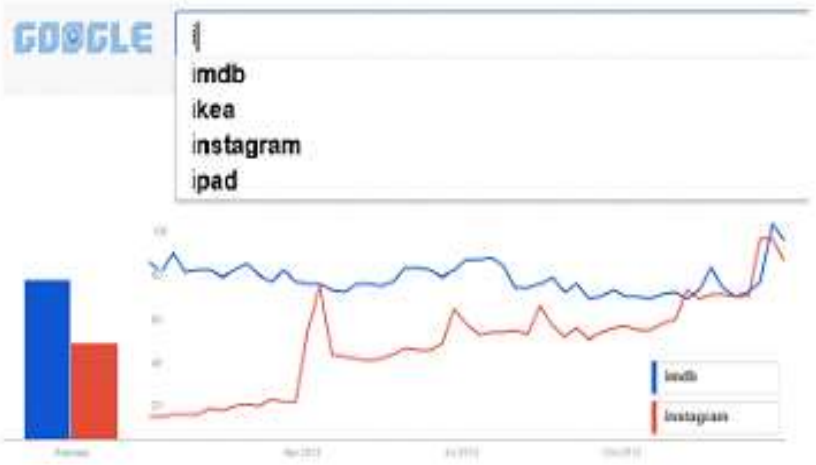


Fig 6. The auto completion results for the prefix i and the snapshot taken in 2013 Jan

Consider the Fig 6 example where the suggestion list for the prefix is shown according to the Google. The imdb is at a higher rank as compared to the Instagram but the popularity of Instagram is rising in the graph. The suggestions depend on the worldwide popularity values but this list can be improvised according to the personalization where the results are modified according to the each individual results. When the user information is unknown then the worldwide results are most appropriate and also the generalized one.

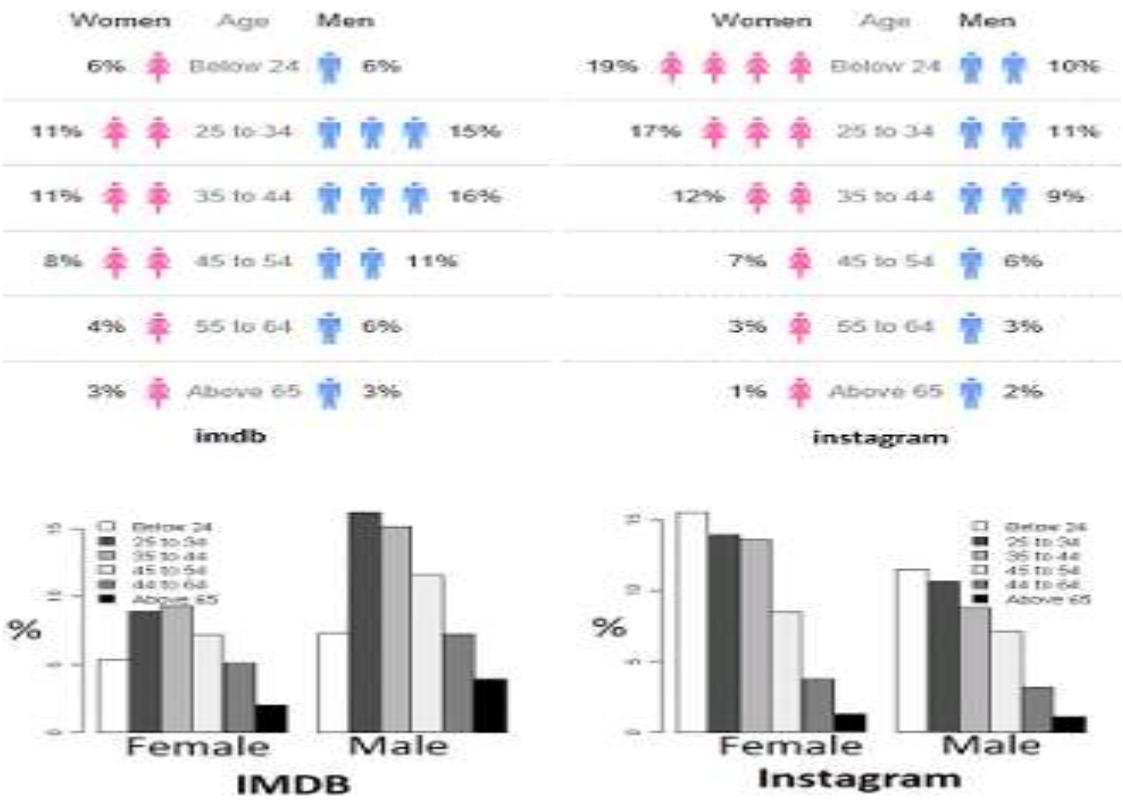


Fig 7. The search results according to the different user and different age groups

Fig 7 depicts the query search of imdb and Instagram depending on the male and female user according to the different age groups. As we can see that the Instagram search by the female users below the age of 44 is more whereas for the same group imdb is more searched by the male users. So if in the Fig 6 query results we incorporate the factor of male and female then the query suggestion can be improvised according to the user. If it is known that the search is done by a female user then Instagram will be at higher rank as compared to the imdb which is vice versa in the case of male users [14].

Chapter 4

PROPOSED WORK

4.1 PROPOSED SYSTEM

Query Auto-completion predictions for the users is a technique to reduce the number of key strokes and recommends about what the user is intended to write. In the writing, there are wide range of techniques e.g. geographical location, context, temporal and seasonal query for the query formulation in the search engine bar. Amongst the most prevalent techniques is the Personalized query results which integrates the user interest areas in the suggestions for the better recommendation. So the user interests can be taken from any of their social networking portals like Facebook like pages or the followed pages in the Instagram.

In this proposed work, Personalized results have the highest considerations then the Context sensitive results are taken into account and then the trending topics related to the prefix keywords are considered.

Initially, user enters the prefix as keywords in the search engine bar. The database at first fetches the trending topics and then extracts the queries which are starting with the current prefix. The trending topics are the Most Popular Completion results which are taken. These suggestions are then entered into a temporary table which is created for every user. The temporary table is created every time user starts entering the keywords in the search engine bar and it gets deleted after suggesting the results.

After the suggestions related to the prefix from trending topics are moved into the table with their count values. Then it is checked if the user is logged in from their FB account or not. An API is developed for accessing the user database with their permission. The permission is taken to use their database to access their likes for knowing the users various interests of the user for the Personalization of the results recommendations. Then their likes database is accessed if it contains any of the query starting with the prefix value are fetched and they are also fed into the

temporary table with their count values as doubled because these results are more preferable and their weightage is considered higher than the trending topics.

After these results are considered then the Context i.e. the queries starting with the prefixes are also evaluated. If any query starting with the keywords is done in the past hour or half an hour are taken. The queries are adjusted according to the keywords position in the query. If the query is starting with the keywords then the suggestions are multiplied with the average value by 4. If the keywords are in between the query then the count value is averaged by 2 and if the query is ending with the keywords then their count value is averaged.

All the results are combined with the count values in the temporary table and they are adjusted in the descending count value. The top 10 results of the temporary table which are predicted as the most possible recommendations for the entered keywords are presented.

4.2 Proposed Algorithm:

The Query Autocompletion technique for the query suggestions

Input: user keywords as prefix P, p ϵ MPC, q ϵ Context, r ϵ Personalized,

ttemporaray table – t, time interval- 1hour,

Personalization- FB likes,MPC- Trending topics, Suggestion - S

Output: Rank list of top 10 query completions for the prefix P

1. for each p ϵ MPC
   for all p , if p ϵ P
   then
   extract the p and count values in t
   end for
2. for each q ϵ Context
   for all q in interval (1 hour),
   if q ϵ P
   then
   extract the q and count values in t
   end for
3. for each r ϵ Personalized
   for all r , if r ϵ P
   then
   extract the r and count values in t
   end for
4. for each S ϵ t
   return the S with their count values in descending order.

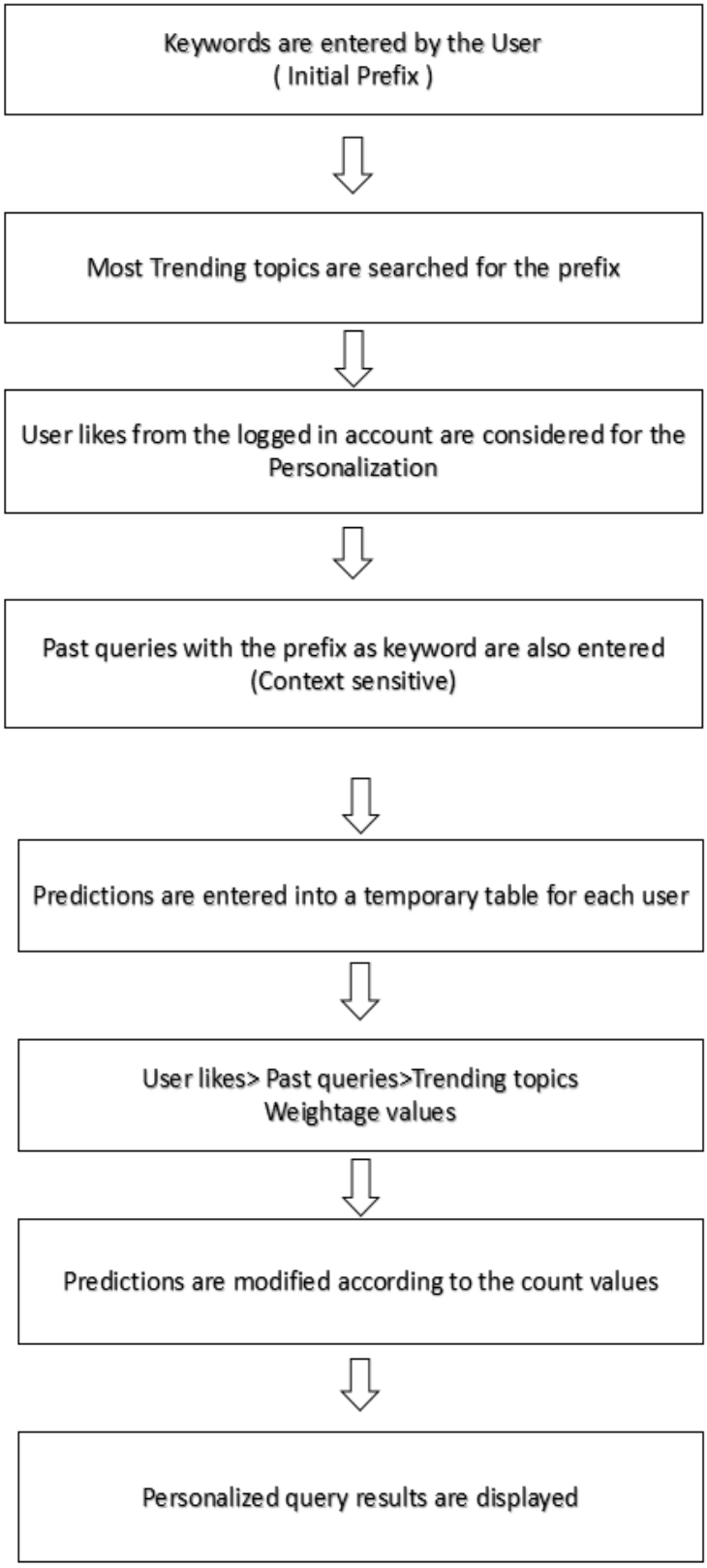


Fig 8. Flow Diagram of Proposed Algorithm

4.3 Prediction evaluation metrics

To assess the result data set of the Query predictions are done on the Hybrid Score for the results. The Hybrid Score is computed on the behalf of the three techniques score value combination.

$$\text{H.Score} = \text{MPC(p)} + \gamma \,.\, \text{CSscore(q)} + (1\text{-}\gamma) \,.\, \text{PSscore (r)} \quad (1)$$

Where $PS_{score(q_c),}$ is the Personalized Score, $TS_{score(q_c),}$ is the Context Value whereas MPC(p) is the Most Popular Score.

The Most Popular Completion results are

$$MPC(p) = \arg\max_{q \in C(p)} w(q), w(q) = \frac{f(q)}{\sum_{i \in Q} f(i)} \quad (2)$$

Where $f(q)$, denotes the number of occurrences of query q in search log Q, and $C(p)$ is a set of query completions that start with prefix p.

The Context Sensitive Score is calculated as,

$$TS_{score(q_c)} \leftarrow \frac{\tilde{y}_{t_0+1}(q_c, \lambda) - \mu_T}{\sigma_T} \quad (3)$$

Where $\sigma_{T,}$ and $\mu_{T,}$ are the mean and standard deviation of predicted popularity of queries in S(p).

The Personalized Value for the prefix p is calculated as,

$$P_{score(q_c)} \leftarrow \frac{P_{score(q_c)} - \mu_P}{\sigma_P} \quad (4)$$

$$P_{score(q_c)} = \omega \cdot Score(Q_s, q_c) + (1 - \omega) \cdot Score(Q_u, q_c)$$

(5)

Where σ_p and μ_P are the mean and standard deviation of similarity scores in S(p).

Personalized QAC works by scoring completions $q_c \in S(p)$ using a combination of two similarity scores Score (Q_s, q_c) and Score (Q_u, q_c) where Q_s relates to the recent likes and Q_u relates to the queries issued before in the session.

Chapter 5

RESULTS

5.1 DATA SET

The experiment for the proposed algorithm is done on the AOL data set which is released in 2006. It contains a large number of users' data and their queries.

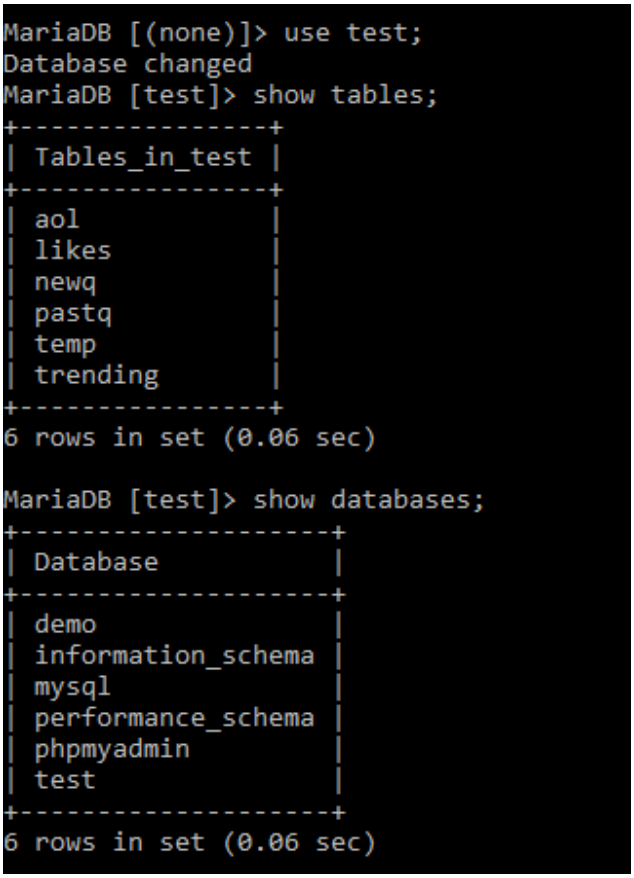


Fig 9. Tables and Database used in Algorithm

Fig 9 contains the databases and the tables which are used in database. The Five tables contains the data set information. The AOL contains the query log of the users for a three month time period. The Likes table contains the users' personal details of FB likes. The pastq table contains the users' context in the previous hour. The trending table contains the most trending topics in

India at that point of time and temp table is created for every user and dropped when the suggestion list is predicted.

```
MariaDB [test]> select query from aol where query like 'ab%' limit 20;
+----------------------------+
| query                      |
+----------------------------+
| abba band                  |
| abba band country          |
| abba band country          |
| abba band country          |
| abba band from             |
| abba band from             |
| abba band bio              |
| abc                        |
| abc 7                      |
| abc                        |
| abercrombie                |
| abercrombie                |
| abercrombie                |
| abcnews.com                |
| abdominal pain and diarhea |
| abdominal pain and diarrhea|
| abdominal pain and diarrhea|
| abctech.com                |
| abraham ramos              |
| abraham ramos              |
+----------------------------+
20 rows in set (0.23 sec)
```

Fig 10. Query set from AOL starting with 'ab' as prefix

```
MariaDB [test]> select query from aol where query like '%ab%'limit 10;
+----------------------------+
| query                      |
+----------------------------+
| www.tabiecummings.com      |
| www.mecab.org              |
| www.acevedoarabians.com    |
| security search and abstract|
| www.juan grabrielcd.com    |
| www.juangabrielcd.com      |
| babiesrus                  |
| babycenter.com             |
| baby center                |
| baby names                 |
+----------------------------+
10 rows in set (0.00 sec)
```

Fig 11. Query set from AOL containing ‘ab’ as keyword in the middle of the query completions

```
MariaDB [test]> select query from aol where query like '%ab'limit 15;
+---------------------------------+
| query                           |
+---------------------------------+
| chocolate lab                   |
| disturbed tab                   |
| stellar kart guitar tab         |
| stellar kart guitar tab         |
| scottie doesn't know guitar tab |
| scottie doesn't know guitar tab |
| 32 below guitar tab             |
| big and rich guitar tab         |
| saab                            |
| indias hijab                    |
| roger schwab                    |
| roger schwab                    |
| roger schwab                    |
| roger schwab                    |
| greyhound rescue and rehab      |
+---------------------------------+
15 rows in set (0.11 sec)
```

Fig 12. Query set from AOL ending with 'ab' as keywords

```
MariaDB [test]> select query from pastq where query like 'ab%';
Empty set (0.44 sec)

MariaDB [test]> select query from trending where query like 'ab%';
Empty set (0.05 sec)
```

Fig 13. Initial query set in the pastq table and trending table

Fig 13 shows that, initially when no user is logged in or no query is done in the previous hour then the pastq and trending table contains the empty data set

5.2 EXPERIMENTAL RESULTS

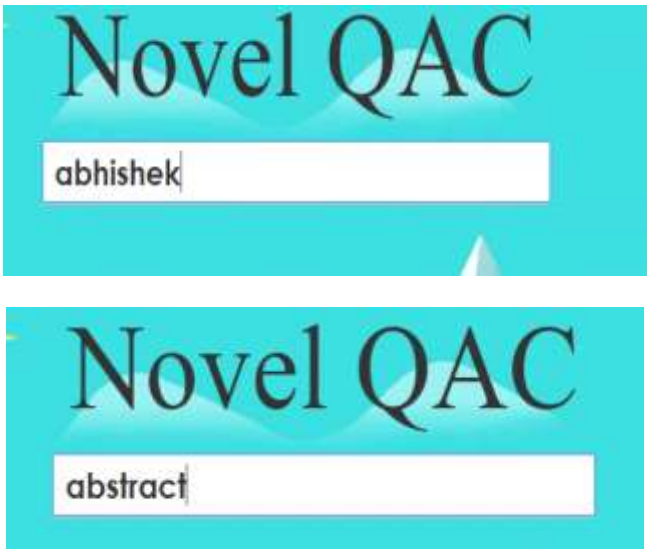


Fig 14. When the queries abhishek and abstract are entered

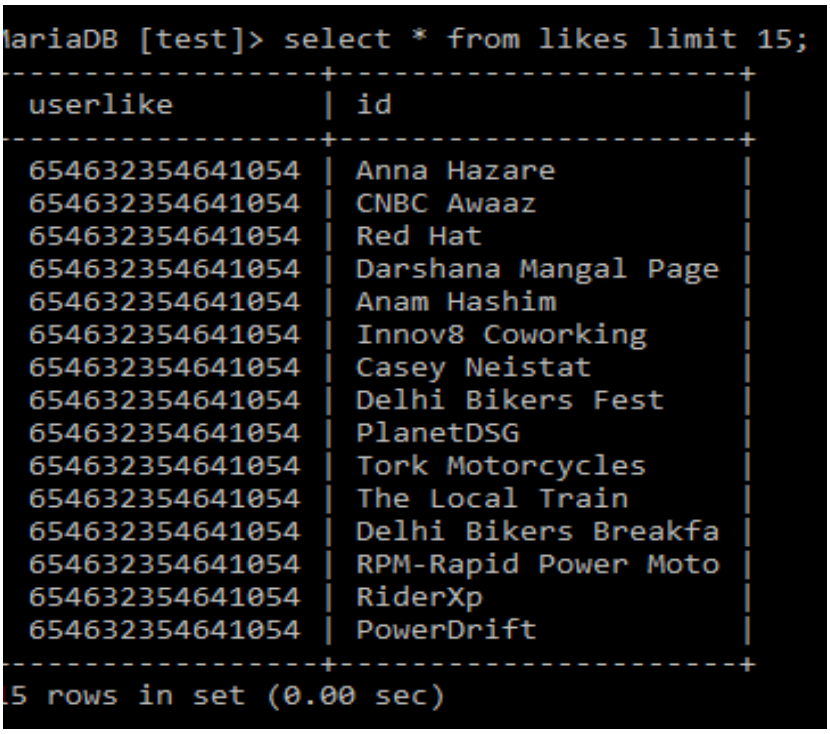
```
MariaDB [test]> select query from pastq limit 15;
+-------------------------+
| query                   |
+-------------------------+
| india                   |
| indian culture          |
| indiana                 |
| friends                 |
| country                 |
| friends                 |
| oval                    |
| city                    |
| cineplex                |
| cinema                  |
| cinematography          |
| how to write a novel    |
| how to dance            |
| how to download a song  |
| what is wikipedia       |
+-------------------------+
15 rows in set (0.48 sec)

MariaDB [test]> select query from pastq where query like 'ab%';
+----------+
| query    |
+----------+
| abhishek |
| abhishek |
| abstract |
+----------+
3 rows in set (0.18 sec)
```

Fig 15. The Pastq data set when the 2 queries are entered with prefix as 'ab'

Fig 14 and Fig 15 shows that when the user enters the queries as abhishek and abstract then they are shown in the pastq table since it is a relevant context for the queries starting with ab.

Fig 16. The Trending topics in India

Fig 16 shows the data set of Most Popular or Trending topics in India at that point of time when the query is done by the user.



```
MariaDB [test]> select * from likes limit 15;
+-----------------+----------------------+
| userlike        | id                   |
+-----------------+----------------------+
| 654632354641054 | Anna Hazare          |
| 654632354641054 | CNBC Awaaz           |
| 654632354641054 | Red Hat              |
| 654632354641054 | Darshana Mangal Page |
| 654632354641054 | Anam Hashim          |
| 654632354641054 | Innov8 Coworking     |
| 654632354641054 | Casey Neistat        |
| 654632354641054 | Delhi Bikers Fest    |
| 654632354641054 | PlanetDSG            |
| 654632354641054 | Tork Motorcycles     |
| 654632354641054 | The Local Train      |
| 654632354641054 | Delhi Bikers Breakfa |
| 654632354641054 | RPM-Rapid Power Moto |
| 654632354641054 | RiderXp              |
| 654632354641054 | PowerDrift           |
+-----------------+----------------------+
15 rows in set (0.00 sec)
```

Fig 17. The user Likes data from Facebook account

Fig 17 shows that when the user is logged in from the FB account then an id for every different user and their likes are stored in the likes table.

```
MariaDB [test]> select * from temp limit 25;
+--------------------+-------+
| Query              | Count |
+--------------------+-------+
| airtel wifi plans  | 86687 |
| air conditioners   | 64565 |
| american idol      |  6030 |
| ask jeeves         |  4977 |
| ask.com            |  3652 |
| amazon.com         |  2361 |
| amazon             |  2051 |
| american airlines  |  1731 |
| ask                |  1348 |
| adserver.ign.com   |  1297 |
| ampland            |  1056 |
| askjeeves          |  1021 |
| airlines           |   930 |
| american express   |   899 |
| aol.com            |   865 |
| aa.com             |   851 |
| amtrak             |   850 |
| a                  |   815 |
| ads.admonitor.net  |   775 |
| adam4adam          |   774 |
| ask jeeves.com     |   735 |
| airline tickets    |   694 |
| ako                |   692 |
| american idol      |   686 |
| autotrader         |   672 |
+--------------------+-------+
25 rows in set (0.00 sec)
```

Fig 18. The Resultant Data in the Temp table

Fig 18 contains the resultant data set for the user starting with prefix 'a' which are combined from all the likes trending and pastq table.

Chapter 6

CONCLUSION AND FUTURE WORK

Query suggestions formulation using the Hybrid Completion technique tends to be very reliable. This can be further improvised by incorporating the present researches together and including new parameters that might be useful. At the same time the speed and accuracy of the suggestions needs to be maintained depending on the dataset we are working on. Thus QAC is a challenge in itself and needs to be researched into. The more involvement of the new parameters will be more useful and more precise results can be predicted.

Some promising future work directions can also be included for the improvement in results. Apart from the three methods presented before, we aim to improve the final method by merging the methods together into one algorithm and implement some new parameters too.

The parameters to be considered are:

- **Sites visited:** The user's browsing history is a reliable indication of the future searches he might perform. Thus by acquiring the user's consent we can mine into the history from browsers to suggest more relevant query completions to the user.
- **Interests/Keywords:** Each user has a certain interests and he/she tends to search more terms related to these interests. Thus we can formulate/collect interests for each user, the algorithm can be further improved.

REFERENCES

[1] Fei Cai, Shangsong Liang, Maarten de Rijke, "Time-sensitive Personalized Query Auto-Completion", in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ser.CIKM'14, New York, USA, ACM 2014, pp.1599-1608.

[2] Surajit Chaudhuri, Raghav Kaushik, "Extending Autocompletion To Tolerate Errors", in Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, ser.SIGMOD '09, Rhode Island, USA — June 29 - July 02, ACM 2009 , pp. 707-718.

[3] Ziv Bar-Yossef, Naama Kraus, "Context-Sensitive Query Auto-Completion", in Proceedings of the 20th international conference on World wide web, WWW '11, Hyderabad, India — March 28 - April 01, ACM 2011, pp.107-116.

[4] Milad Shokouhi, Kira Radinsky, "Time-Sensitive Query Auto-Completion", in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12 , USA — August 12 - 16, ACM 2012 , pp.601-610.

[5] Fei Cai, Shangsong Liang, Maarten de Rijke, "Prefix-Adaptive and Time-Sensitive Personalized Query Auto Completion", IEEE Transactions on Knowledge and Data Engineering, Vol. 28 , no. 9 , IEEE 2016, pp. 2452-2466.

[6] Ryen W. White, Gary Marchionini, "Examining the Effectiveness of Real-Time Query Expansion", in Information Processing and Management: an International Journal archive, Vol. 43, no. 3, May 2007, NY, USA, pp. 685-704.

[7] Nadav Golbandi Golbandi, Liran Katzir Katzir, Yehuda Koren, Ronny Lempel , "Expediting Search Trend Detection via Prediction of Query Counts", in Proceedings of the sixth ACM international conference on Web search and data mining, WSDM '13 ,Rome, Italy — February 04 - 08, ACM 2013 , pp. 295-304.

[8] Taiki Miyanishi, Tetsuya Sakai, "Time-aware Structured Query Suggestion", in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '13 ,Dublin, Ireland — July 28 - August 01, ACM 2013 , pp. 809-812.

[9] Milad Shokouhi, "Detecting Seasonal Queries by Time-Series Analysis", in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11 , Beijing, China — July 24 - 28, ACM 2011, pp. 1171-1172

[10] Alisa Strizhevskaya, Alexey Baytin, Irina Galinskaya, Pavel Serdyukov , "Actualization of Query Suggestions using Query Logs", in Companion Proceedings of the 21st International Conference on World Wide Web, WWW '12 , France — April 16 - 20, 2012 , pp. 611-612.

[11] Anagha Kulkarni, Jaime Teevan, Krysta M. Svore, Susan T. Dumais, "Understanding Temporal Query Dynamics", in Proceedings of the fourth ACM international conference on Web search and data mining, China — February 09 - 12, ACM 2011 , pp. 167-176.

[12] Steffen Bickel, Peter Haider, Tobias Scheffer, "Learning to Complete Sentences", in Proceedings of the 16th European conference on Machine Learning, ECML'05, Portugal — October 03 - 07, 2005, pp. 497-504.

[13] Rodrygo L. T. Santos , Craig Macdonald, Iadh Ounis, "Learning to rank query suggestions for adhoc and diversity search", in Journal of Information Retrieval archive, Vol. 16 , no. 4, August 2013, pp. 429-451.

[14] Milad Shokouhi, "Learning to Personalize Query Auto-Completion", in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '13 , Ireland — July 28 - August 01, ACM 2013 , pp. 103-112.

[15] Zhen Liao, Daxin Jiang, Enhong Chen, Jian Pei, Huanhuan Cao, Hang Li , "Mining Concept Sequences from Large-Scale Search Logs for Context-Aware Query

Suggestion", in ACM Transactions on Intelligent Systems and Technology (TIST) archive, Vol. 3, no.17, ACM 2011.

[16] Qingqing Gan, Josh Attenberg, Alexander Markowetz, Torsten Suel, "Analysis of Geographic Queries in a Search Engine Log", in Proceedings of the first international workshop on Location and the web, LOCWEB '08 , , China — April 22 - 22, ACM 2008, pp.49-56.