Learning techniques for Query Auto-Completion and Optimising Search Engine Results

Mohammed Kamran
Department of Computer Engineering
Aligarh Muslim University
Aligarh, India
kamranisg@gmail.com

Abstract - **We live in a society where technology is growing at a rapid rate. The idea of having access to knowledge right at our fingertips was a mere dream 30 years back. With the advent of Internet and web search engines, our view of world has changed completely. We require smarter search engines employing algorithms which can enhance user's search experience. This paper throws light on Auto-completion techniques and Query optimisation methods which can help users for better and faster results. In the past the suggestions of user related queries would depend on past query log data which was called the traditional batch learning method. However, this method didn't hold good when queries were dynamic and carried weights of importance. Researchers have come up with techniques that help users get their required results in a more adaptive manner.**

Index Terms - Query auto-completion, Multi-Armed Bandits, Query Optimisation, Exploitation vs Exploration

1. INTRODUCTION

Web search engines provide a gateway for information to flow from websites to users. They contain an input dialog box whereby a user types his query.

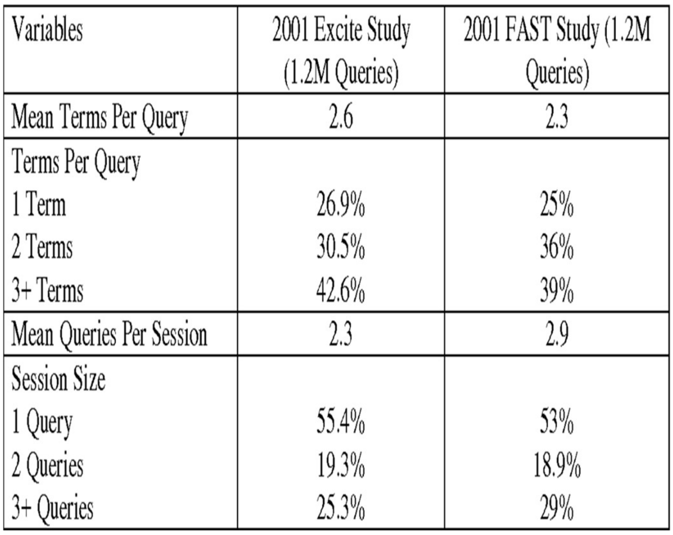
Due to enormous searches on web, it becomes the need of hour that we rank each web search query and suggest relevant queries whenever a user starts typing in input dialog box. Each web search engine employs its own algorithms to rank queries, auto-complete and optimise them.

In a 2001 research study conducted on Excite web search engine revealed important and interesting facts of web search engine as shown in Table 1.

About 26.9% of users entered only a single keyword while 3 or more words were entered by 42.6% of users. These results brought the mean query length to be around 2.6 terms. The same statistics can be viewed from FAST search engine from Table 1.

More over about 50% users just clicked on top two websites displayed on the screen. The study concluded with the fact that query distribution follows power law where shorter terms enjoys larger number of queries typically around 100 million queries. Queries which have longer phrases such " Which is the nearest restaurant? " requires the use of Natural Language processing. However, we limit our study on shorter phrases in this paper.

Table 1.
Performance of Excite Search Engine



Variables	2001 Excite Study (1.2M Queries)	2001 FAST Study (1.2M Queries)
Mean Terms Per Query	2.6	2.3
Terms Per Query		
1 Term	26.9%	25%
2 Terms	30.5%	36%
3+ Terms	42.6%	39%
Mean Queries Per Session	2.3	2.9
Session Size		
1 Query	55.4%	53%
2 Queries	19.3%	18.9%
3+ Queries	25.3%	29%

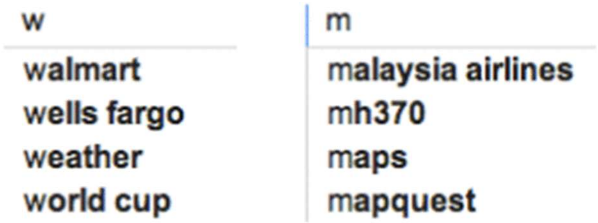
Researchers have come up with Query Auto-completion (QAC) system and Query optimisation methods to reduce users' efforts in typing. This technology is helpful in portable devices such as Mobile phones which require extra effort in typing words due to limited space.

QAC system rather than predicting on the basis of individual candidates past queries, uses the results of all candidates and rank them accordingly.

This method is called Most Popular Completion (MPC). MPC system is dynamic and runs in an online mode. However, aggregating past queries in this model would imply serious implication such as in, time-dependent events and unpredictable events.

For example, Fig. 2 shows the suggestion list of a user typing a keyword 'r' on 19 April 2019, the result shows rcb vs kkr on top. Due to ongoing Indian Premier League, the most searched query was of the match between Royal Challengers Bangalore (rcb) and Kolkata Knight Riders (kkr) on 19 April 2019.This is classic example of time dependent event. We want web search engines to reflect this change timely.

Similarly, the day after the crash of Malaysian Airlines MH17 the following results were displayed as shown in Fig. 1. These strategies have a disadvantage of being stuck in a local minimum. Thus, QAC system would employ an online decision-making problem.



w
walmart
wells fargo
weather
world cup

m
malaysia airlines
mh370
maps
mapquest

Fig 1. Suggestion lists of a major search engine. Left: Prefix *w*, screenshot taken on July 12, 2014. Right: Prefix *m*, screenshot taken on July 18, 2014.

2. RELATED WORK

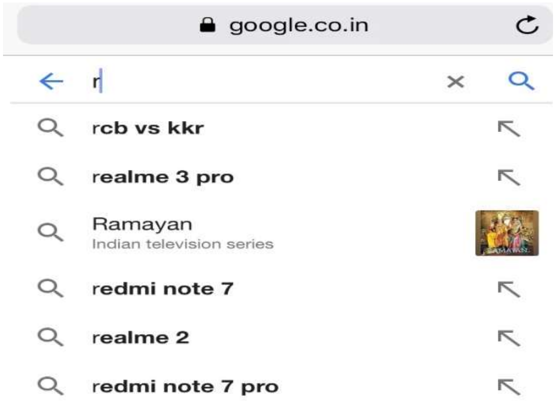
2.1 Query Auto Completion

2.1.1 MPC In QAC System

QAC techniques help users by predicting queries beforehand on their search engines. MostPopularCompletion(MPC) is a type of QAC algorithm which gives frequency scores on user's past popularity from query database . Kraus has taken into account the similarity between QAC users by proposing a QAC sensitive based model of prediction. An improved MPC technique was suggested by Radinsky which required the use of recursive time series models.

2.1.2 Google Autocomplete API

Another very useful work is of Google, which has developed a Query Autocomplete API. This API is useful in finding geographic locations based on your query. A user who types "pizza near San Francisco", will get the all the restaurants which serves pizza near San Francisco. The Autocomplete request is a URL of the form "https://maps.googleapis.com/maps/api/place/queryautocomplete/output?parameters"



google.co.in

r

rcb vs kkr
realme 3 pro
Ramayan
Indian television series
redmi note 7
realme 2
redmi note 7 pro

Fig 2. Character 'r' typed on Google

2.1.3 Ranked MAB for QAC

The main objective to tackle uncertainty in decision making would be to maximise cumulative rewards using existing knowledge and acquire new knowledge through actions. This is called the fundamental dilemma of exploitation and exploration. This leads us to the classical problem of reinforcement learning called Multi arm bandits (MAB). Here we aim to maximise rewards by distributing among possible choices. The name comes from the idea where a gambler is given a set of slot machines who decides which machine to start with, how many times to start the machine and in which order to start. The bandit problem has several practical applications in clinics to minimise patient losses and in networks to reduce delays and cost.

The Multi-Armed Bandits approach addresses the problem of exploitation vs exploration. It exploits well played query results and explores less relevant query results so that they perform well in future use.

Consider t, to be number of users participating in an online fashion and j be the number of characters stroked on keyboard by **tth** user. At time **(t,j)** possible queries set is formed called $\mathbf{A_{(t,j)}}$. The QAC system proposes a suggestion list of n queries such as

$$Q^{x}_{(t,j)} = (q^{1}_{(t,j)},q^{2}_{(t,j)},.....,q^{n}_{(t,j)}). \qquad (1)$$

Each user will click at most one query. This will help QAC system to get rewards after each j^{th} keystroke if user picks up any query from the suggested list presented before him/her. This ranking in Multi-Armed Bandits helps in improving user experience and simultaneously utilises the less preferred queries by pushing out of local minima.

2.1.4 QAC by aggregating Web search engines

Another possibility that researchers suggest is to keep one search engine as a mixture of many engines that would be at par with all of them. The objective here is to maximise number of user clicks on suggestions provided in the list and bring out top k-relevant queries on the list. Each query $q_{(i)}$ is given position $p_{(i)}$ from a list of m queries. A total of m search engines are allocated to each query $q_{(i)}$. Each search engine $S_{(i)}$ is responsible for position i . The goal of this aggregation is to choose the best search engine at each position i .

Researchers have come up with 4 different algorithms namely The Ranked Bandits algorithm for QAC-ME , Cascade Bandits algorithm for QAC-ME , Explicit Ranked Bandits for QAC-ME and Explicit Cascade Bandits for QAC-ME.Bandit based aggregation algorithms have proven to reliable , fast and accurate techniques in choosing web search engines smartly.

The future of Search Engine Autocompletion rely on algorithms that are time and space efficient. This growing field is in need of studies which can bring a dynamic change in the era of Query optimisation in Web Search engines.

2.2 Query Optimisation

The objective is to produce results that are parallel with what user's intent on having from the search engines. A major difficulty arises when users are not able to find relevant results. They have to search through all sites to get their information. This type of problem is typically referred to as Information Kill Problem

2.2.1 Mining query logs

Suggesting queries from document-based data and log-based data are two categories of studies. The latter method is a more popular option. Query logs have been interest over the years and a lot of research has been undergoing to bring out much better results out to user.

A typical Azure log contains the following attributes in the database: (A) Time Generated, (B) Computer used, (C) EventLevelName , (D) Type of query , (E) SourceSysten , (F) Source .

2.2.2 Exploiting Query Similarity in keywords

Two queries are said to be more similar if they contain identical keywords. Consequently, the formula proposed is as follows;

$$s1^{(x,y)} = c^{(x,y)}/(\max(w^{(x)},w^{(y)}) \qquad (2)$$

Where $w^{(x)}$ and $w^{(y)}$ denote the number of keywords in queries x and y respectively, $c^{(x,y)}$ represent number of common words between x and y , $s^{(x,y)}$ represent the similarity measure between x and y.

From the above formula, we can perceive longer queries to be giving better results. Studies of Radecki and Li and Danzig proposed similarity between canonical Boolean terms.

2.2.3 Exploiting Similarity based on URL clicking

Another measure to check similarity between two queries is based on URL clicking. If two URL clicks of user result in same document then, those two are more similar. Consequently, the formula proposed is as follows;

$$s2^{(x,y)} = com^{(x,y)}/(\max(ref^{(x)},ref^{(y)}) \qquad (3)$$

Where $ref^{(x)}$ and $ref^{(y)}$ denote the number of documents referred for queries x and y respectively, $com^{(x,y)}$ represent number of common documents clicked between x and y , $s^{(x,y)}$ represent the similarity measure between x and y.

3. CONCLUSION

The subject of Query Auto-completion has been an interest to researchers all over the globe since the advent of internet and the search engines. The studies presented above have overcome the traditional batch learning methods and have come up with more reliable approach using the Multi-Armed Bandits in QAC system. Further work using Thompson sampling in QAC is also been done. The Query Auto-complete API of Google is great feature of a what a modern search engine must have. Furthermore query log mining has been done to optimise queries in full measure. Overall, it can be said that since users are growing, so is the data. This data is of various forms and in huge volumes. Gone are the days when oil used to be the currency of the world. This world is in an ocean of big data and we need algorithms to exploit to its fullest.

REFERENCES

[1] Y.Wang, H.Ouyang,H. Deng, Y. Chang, "Learning Online Trends for Query Auto-Completion" in IEEE Transaction on Knowledge and Data Engineering,2017,pp 2442-2454.

[2] Jiawei Liu, Qinqshan Li, Yishuai Lin, Yingjian Li, "A Query suggestion method based on random walk and topic concepts" in IEEE/ACIS 16th International Conference on Computer and Information Science,2017