User Web Search Behaviour

Murtuza Kutub
murtuzafirst@gmail.com

Prachetaa R
prachetaa@gmail.com

Mangesh Bedekar
bedekar@bits-goa.ac.in

Birla Institute of Technology and Science – Pilani, K. K. Birla, Goa Campus,
NH-17B, Airport Road, Zuari Nagar, Goa, INDIA. PIN – 403726.

***Abstract*---Since the inception of the web searching technology, people have been searching for almost everything and anything on the internet. The ever-increasing dependency of users on these search engines and the dynamic nature of the World Wide Web has reduced the accuracy of the search results and increased the search time of an individual. Today, more than ever before, there is a need for search engines to be relevant and precise to the user's needs and to be able to make decisions about what the user wants to search , and should be able to suggest him similar or related topics of his interest. This increasing need of the search engine to become a decision engine [1] (term coined by Stefan Weitz [2]) brought to fore various creative technological ideas like Tag clouds and AutoComplete [3].**

For a better and more relevant search experience, it is crucial that we study the present search behavior of users and its corresponding response by the search engine. This work contemplates the nature of searches made and how they evolve from time to time. In this paper we examine and construe data from various angles and then provide our suggestions and conclusions for a better, more personalized and relevant search.

Keywords -Long tail phenomenon; Personalized search; Tag cloud; Data mining; Auto complete; Hot trends; Search experience; 'Headophile'; 'Tailophile'; Decision engine

I.INTRODUCTION

The World Wide Web is growing exponentially and there is a need for easier and faster ways to search data, i.e. basically making search more user friendly so that the user can find his/her page of interest faster. This calls for the need of better ways of projecting data to the user depending on the individual's interest. More visually appealing ways make it easier for the user to spot the keyword searched for by the user.

Showing data in the form of tag cloud [4] displaying the latest or hot trends [5] of the day, keeping in account the search behavior of the particular user is one way of transforming a search engine into a decision engine, as the search engine makes a decision of what results or suggestions to show for a particular query based on the user's searching history.

Another important area is that of the long tail. The long tail phenomenon [6] has been observed in web searching. The long tail phenomenon, however, is in principle consistent with two fundamentally different hypotheses. The first, and generally accepted theory, is that a majority of consumers prefer popular offerings while only a minority seek niche content; the second hypothesis is that everyone is a bit eccentric, consuming both popular and specialty products.

To distinguish between these possible alternatives and also to suggest better ways of a searching experience, we examine extensive log data on user preferences for Web search and Web browsing. From the results, we find overwhelming evidence that nearly everyone is at least a bit eccentric. Our work highlights the diversity of individual tastes in a student campus and the changing mentality of humans every day.

The remainder of our paper is organized as follows. In section II we review related work. Section III describes our approach, system design, system description and analyze the data. We conclude in Section IV by discussing our results and contributions through this paper.

II.RELATED WORK

Our work in this paper will be suggesting on improving the search experience by exposing the user to his own user specific search environment. The work in the following fields has helped in improving the user's search experience by leaps and bounds.

The long tail phenomenon has been an area of research for many years. This phenomenon has been observed in many fields such as web searching, inventory stores such as Amazon [7] showing the fact that long tail is also an important part from which revenue can be generated. The "long tail" view was coined and popularized by Chris Anderson [8] to describe consumers' demand for niche products in an age of infinite-inventory retailers.

978-0-7695-4246-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ICETET.2010.31

Tag cloud is a spatial distribution of tags (words/keywords) which are displayed using different styles (font colors, font size etc.) based on their individual popularity. Most of the websites today use tag clouds as it helps the user to save time and efforts to look for information he/she is interested in and it is also visually very appealing.

Auto complete [9] is a feature provided by many search engines, web browsers, editors etc. which predicts a word or a phrase that the user wants to type in without the user actually typing it in completely thus helping to speed up the search process. For e.g. Google Auto complete is an interesting and useful feature which suggests the user related queries which he/she might be interested in based on his/her input of initial few letters.

Hot trends are another topic of interest in which the hottest topics in discussion, for a particular day can be viewed. For e.g. Google hot trends [10] implemented by Google is a feature in which a particular day’s top 20 searches are shown for each country. This feature helps in identifying the topics which are of current interest in the world.

III. DATA MINING AND ANALYSIS

A. Our Approach

a. Mining

The web usage of approximately 1200 students in BITS-Pilani, Goa campus [11] was logged between the periods of 7th of March 2010 to 12th of March 2010. The students were separated into 7 residential sectors scattered throughout the whole campus. The main aim was to log the web activity of students which included every possible type of individual in different age groups and genders.

Web surfing is seen as a very useful, interesting, helpful and loads of data is available. The log file on the other hand is actually seen as useless, but when proper data is mined from the log file, loads and loads of information can be gathered from the activity logs.

The log file was generated through the process of logging the proxy server (Squid) [12]. The log file was of the following format and was hash separated.

Table I. Database Format

Date-time	**IP**	**URL**

Each log file had over 12 lakh entries which included web surfing apart from Google searches [13]. The next process was to remove the useless data so as to get useful information from the log file through shell script which removed all the useless URLS and kept only the Google search URLs.

The shell script [14] was preferred over any other language because of its efficiency, simplicity and speed. This step of filtering was done so that the entire process could be made faster and memory efficient.

The most important point to note here is that although the logging was done with the IP, at all points the anonymity of the user was maintained.

The data in the log file was hashed so that the true identity of each user was hidden.

The next step involved importing the data into to MySQL [15] database. The MySQL database had tables with following fields

- Date time which was of date time type.
- IP which was of text type.
- URL which was of long text type.

The log data was imported using PHP [16] code. Once imported, the URLs were parsed to get the search queries which were then imported into another table.

We also found the number of unique search queries and their frequency of occurrence on a hourly and daily basis.

Using this information, keyword specific graphs with date on the x axis and their respective counts on the y axis were plotted. In addition to this, a bar graph of the hourly variation of Google queries for a particular date was also plotted. A date specific tag cloud of Google search queries was also made.

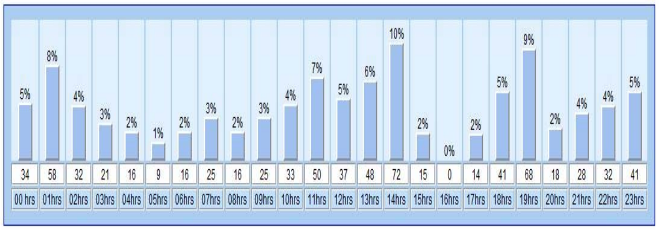


Figure 1. Bar graph showing Google search on an hourly basis.

b. System Design

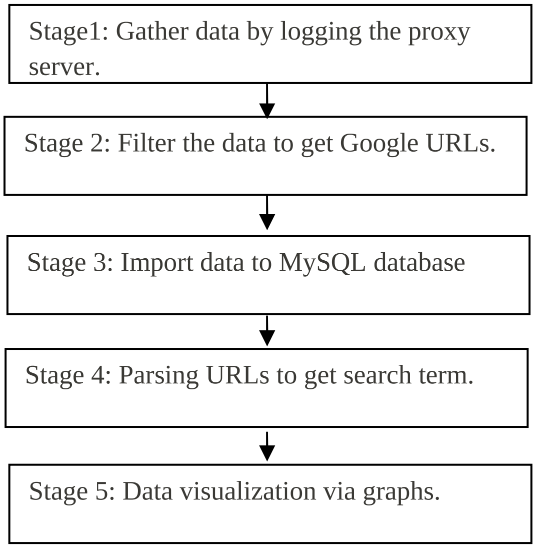

Stage1: Gather data by logging the proxy server.
Stage 2: Filter the data to get Google URLs.
Stage 3: Import data to MySQL database
Stage 4: Parsing URLs to get search term.
Stage 5: Data visualization via graphs.

Figure 2. Block Diagram of our System.

c. System Description

The stages involved in the process are as follows-

Log the web usage of individuals so as to get the web surfing details through the proxy server.

Web activity logging was done to maintain the record of web activity.

The unnecessary URLs were removed at this stage, so as to reduce the wastage of memory and processing time/power in the upcoming stages.

All of the web activity that was not Google search was removed and the remaining was copied to another file. This brought down the processing time in the database from approximately 10 minutes per log file to a couple of seconds.

We then import the URLs to MySQL database with the specified fields present as mentioned above.

This was done for enabling the remaining stages which were done using PHP and MySQL. Data management is easier, faster and also more secure in a database.

The information of interest in the Google URL is the Google search query which was extracted. Duplicated entries were counted and removed preserving their count, date and time.

We then went on to plot graphs and charts so as to visualize data and get useful insights from the activity log.

Results are easy to comprehend when presented in the form of graphs and charts rather than numbers.

B. Data Description

Table II. Descriptive statistics for the dataset analyzed. Observations correspond to the searches made to Google.

Days	**Sunday**	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Total Log Entries	12,74,777	19,01,701	14,17,241	20,21,069	16,17,261	17,79,185
Google Searches	570	1,242	497	435	400	714
Percentage Google Searches	0.045 %	0.065 %	0.035 %	0.022 %	0.024 %	0.040 %

Our empirical results are based on an analysis of user behavior across large datasets obtained from logging of surfing data and extracting only Google searches. Summary statistics for these datasets from the activity logs are given in Table II.

As we are primarily concerned with Google searches only, other activity logs were excluded. The total number of log entries for nearly a week was 2,00,07,029 out of which the total number of Google searches were 7,708. The whole process of sieving the Google URLs out of the entire web activity logs took approximately 30 seconds for each log file which were around 200 to 240 MB’s.

There were differences seen in the total number of searches every day.

There was a sudden rise in the number of searches and also the number of entries in the logged data on the start of the weekday from a mere 570 Google searches to a whopping 1,242 searches showing an increase of more than 100%.

From the activity logs, it can also be seen that the number of hits for a particular search depends on the popularity of the search term on that particular day. We define the popularity of an item (a search term) to be the fraction of total consumption fulfilled by that item.

For e.g. There was a sudden increase in the search term " IPL " and its related terms from NIL search queries on 11th March 2010 and a jump to 20 searches on 12th March 2010. The same style was also seen on Google hot trends wherein the search query 'Michael Jackson' went up leaps and bounds when he breathed his last on 25/06/2010. After a few days the search count went down gradually to NIL on 30/06/2009.No search terms were observed before 24/06/2010.[17]

TABLE III. Google Searches Weekend v/s Weekday.

Sunday	570 Google searches
Monday	1,242 Google searches

C. *Analysis of Data*

From the datasets, it has been observed that from an initial log file which contained the entire web surfing activities (12 lakh entries), the sieved log of only the Google searches had roughly around 550 entries.

This shows that a mere 0.04 % of web usage constitutes searching (Google).

TABLE IV. Percentage of Web traffic v/s Google search traffic.

Total web surfing	**Google searching**
100 %	0.04%

Consistent with past work on the long tail, we find in the datasets that (1) a relatively small number of items account for a disproportionately large fraction of total consumption; and (2) the tail, in aggregate, is nevertheless relatively heavy.

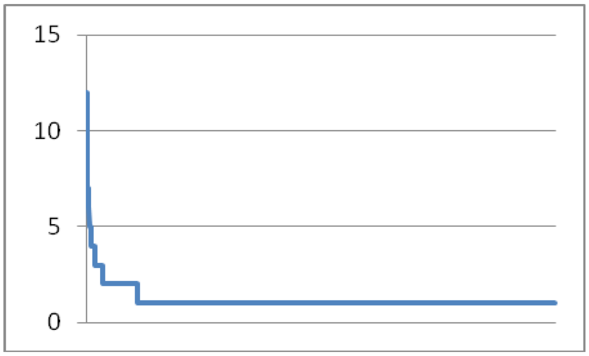


Figure 3 Averaging the six graphs, final graph showing the presence of long tail present in the plot of search term v/s count.

Around 3% of the total search terms constitute of the head and the remaining belong to the tail. This shows that huge revenue can be made out from the 97% tail which Google did in the feature ad words.

The 97% Figure is pretty huge considering the fact that so many unique searches are made which shows the need for a better personalized search engine. Very few people search keywords belonging to the head whereas a huge number of people search keywords belonging to the tail. It is also seen that the long tail keeps evolving with time.

The keyword specific graph with variation of date on the x axis and its respective count on the y axis can be used as a powerful revenue generating tool for commercial search engines to find out the topics that transform from the tail to the head and hence invest in the budding hot topics.

IV. OUR CONTRIBUTION

A. Personalised Search v1.0

Looking at the extensive data on search behavior, we find overwhelming evidence that a vast majority of users are a little bit eccentric, having idiosyncratic tastes. Moreover, our analysis provides an insight on the necessity to capitalize on the long tail also and the need for more refined technologies for specific population and richer representation for crawling by standard web engines.

Given the observed user eccentricity, one reasonable hypothesis is that users consume content proportional to popularity, but otherwise do not differentiate between head and tail items. Finally it has been seen that users generally appreciate the tail less than the head, in turn diminishing its importance.

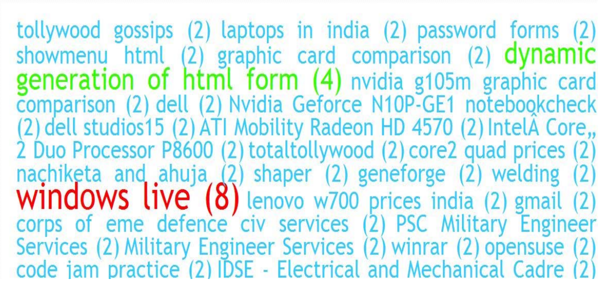
Using the long tail analysis, commercial search engines can venture into the unexplored fields of personalized search. The entire process of Google search query extraction could be narrowed down to individual users from which the individual's search affinity towards the head and the tail can be found out.

Let's say for example **-** When a user logs into his Google account, Google can give him the current hot topics of the day if he is a 'Headophile' (concentric individual).For a 'Tailophile', Google can analyze the search trends and suggest hot topics from the head which match those trends.

'Headophile' - When the majority of a person's searches fall in the head region of the long tail graph, he is said to be a head search loving person, or we call him the 'Headophile'

'Tailophile' - When the majority of a person's searches fall in the tail region of the long tail graph, he is said to be a tail search loving person, or we call him the 'Tailophile'

Since in a college, the search queries of interest maybe similar at a given time, the Auto complete feature will help in making searching easier and faster and helping the individual reach his information faster. The Auto complete feature if made by taking into consideration the group of people whose suggestions he acknowledges and his/her own search behavior, will lead to better auto completion results.



tollywood gossips (2) laptops in india (2) password forms (2) showmenu html (2) graphic card comparison (2) dynamic generation of html form (4) nvidia g105m graphic card comparison (2) dell (2) Nvidia Geforce N10P-GE1 notebookcheck (2) dell studios15 (2) ATI Mobility Radeon HD 4570 (2) IntelÂ Core,, 2 Duo Processor P8600 (2) totaltollywood (2) core2 quad prices (2) nachiketa and ahuja (2) shaper (2) geneforge (2) welding (2) windows live (8) lenovo w700 prices india (2) gmail (2) corps of eme defence civ services (2) PSC Military Engineer Services (2) Military Engineer Services (2) winrar (2) opensuse (2) code jam practice (2) IDSE - Electrical and Mechanical Cadre (2)

Figure 4. Tag cloud showing hot Google queries.

In addition to this, a display of a tag cloud for a given date will also help us in finding what is the most currently searched topic in a day and will help in reducing the bandwidth required as the hottest query page can be cached so that the user is directly presented with the page rather that the results of the search thereby making the whole process faster.

B. Hourly graph – Resource Management

The hourly graph is drawn taking into consideration a 24 hour log as mentioned previously. The major benefit of doing an hourly analysis is that on the server side, resources could be managed accordingly.

For e.g. Since our work deals only with Google searches, we could suggest the search giants, Google to allocate more bandwidth during the peak hours which can be found out from the hourly plot and remove the surplus bandwidth present during the 'low hours'.

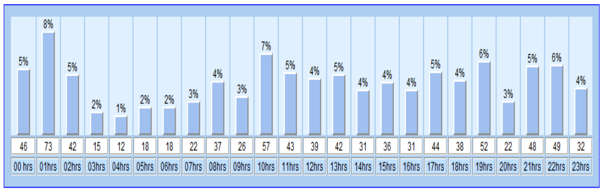


Figure 5. Hourly graph of student residential sector on a weekday

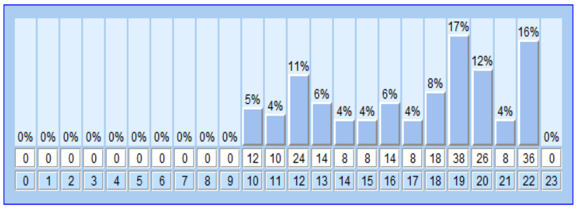


Figure 6. Hourly graph of institute on a weekday

On a campus level, since total bandwidth is divided among different sections like faculty, institute, student residential sectors etc., using the graph, we see that the major percentage of searches done in the student residential sectors is from 21:00 hrs to 03:00 hrs (approx. 30 %). Since during these times the resources allocated to the institute is idle, this can be redirected to the student residential sectors and the faculty. This way, an **intelligent** and **efficient** management of resources is possible.

C. Long tail – the 'cut'

The still unanswered question about where to cut the long tail, holds the potential to generate huge amount of revenue for the search engine. For eg. Google exploited this long tail with the concept of 'Adwords' which is now printing billions of dollars for them.

Finding where to cut the long tail is basically balancing the equation of maximizing profits and minimizing resource usage. The head region constitutes a fraction of search queries and its Gain per unit Resource (GpR) used is much higher. Since the long tail almost tends to infinity, the cut has to be made at a point beyond which the GpR is considered insignificant for the respective search engine.

D. Meta – Data

Here are some insights into the data collected.

Prior to any analysis, data cleaning was done, which involved manually removing erroneous searches which was copy pasted from somewhere. The entire data was scanned through to remove such junk queries, out of which no valuable information could be earned.

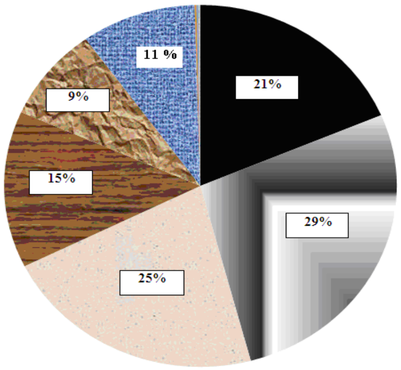


Figure 7. Pie chart showing the percentage of word lengths.

Out of the remaining data available, it was observed that the average search query character length was calculated to be approximately 17 characters. The cumulative maximum was 84 and the cumulative minimum was 2.

This reveals that not many individuals search for words of length more than 17 characters on an average.

The average search query word length was between 2 and 3 words with the cumulative maximum being 13 and minimum being 1.

This shows that very few users search keywords of length more than 3 words on an average.

It was observed that on an average 89 % of the search queries were between 1 and 5 words inclusive and the rest belonging to above 6 words.

One more interesting fact to note here is Google's auto complete feature. It has been observed that Google's auto complete does not suggest any search query which is more than 5 words in length.

TABLE V. Query word length v/s search volume.

Query Word Length	Percentage of search volume
1	21
2	29
3	25
4	15
5	9
1-5	89
6-18	11

REFERENCES

[1] and [2] Available at - http://searchengineland.com/bings-stefan-weitz-rethinking-the-search-experience-34165

[3] Available at - http://en.wikipedia.org/wiki/Autocomplete

[4] Available at - http://en.wikipedia.org/wiki/Tag_cloud

[5] Available at - http://www.google.com/trends/hottrends

[6] Available at - Chris Anderson. The long tail. Wired Magazine, 12(10):170{177, 2004.

[7] Available at - http://www.amazon.com/

[8] Chris Anderson. The Long Tail: Why the Future of Business is Selling Less of More. Hyperion, 2006.

[9] Available at - http://code.google.com/p/jquery-autocomplete/

[10] Available at - http://www.google.com/trends/hottrends

[11] Available at - http://www.bits-goa.ac.in/

[12] Available at - http://www.squid-cache.org/

[13] Available at - http://www.google.co.in/

[14] Available at - http://en.wikipedia.org/wiki/Shell_script

[15] Available at - http://www.mysql.com/

[16] Available at - http://www.php.net/

[17] Available at - http://www.google.com/trends/hottrends?sa=X&date=2009-6-24

http://www.google.com/trends/hottrends?sa=X&date=2009-6-25

http://www.google.com/trends/hottrends?sa=X&date=2009-6-26

http://www.google.com/trends/hottrends?sa=X&date=2009-6-27