CHAPTER 3

USERS SEARCH TRENDS ON WWW AND THEIR ANALYSIS

3.1 CURRENT STATISTICS OF WWW

Approximately 1.67 billion of world population (world population approx 6.77 billion) use Internet [100] (see Figure 3.1) i.e. 24.7% of the total population (see Figure 3.2). In fact from .36 billion in 2000, the number of Internet users has increased to 1.67 billion in 2009 i.e. an increase of 362% from 2000 to 2009 (see Figure 3.3). Figure 3.4 shows the Internet user’s growth in the world by regions. Same growth rate is expected in the future too. Geographic region wise percentage of population using Internet in the World is as shown in Figure 3.5. From this figure, it may be concluded that North America is at the top, where approximately 70% of the population use Internet followed by Oceania/Australia and Europe respectively. Asia is at the second last position, only a little ahead of Africa, where 18.5% of population uses Internet. As far as, the total numbers of user’s world wide are concerned, Asia is at the top with 42% Internet users from this sub continent followed by Europe and North America (see Figure 3.6). India is the third largest country of Internet users after China and Japan in Asia, where approximately 81 million people use Internet as shown in Figure 3.7.

![Figure 3.1: Internet Users in the World by Geographic Regions](image-url)
**Figure 3.2:** Percent of Population Using Internet in Asia and Rest of the World

**Figure 3.3:** Internet User Growth (2000-2009) in Asia and Rest of the World

**Figure 3.4:** World Internet Users Growth (2000-2009) by World Regions
Figure 3.5: Percentage of Population Using Internet in the World by Geographic Regions

Figure 3.6: World Internet Users by World Regions

Figure 3.7: Top 10 Internet Users in Asia (Total 704 million approx.)
From the statistics discussed, it may be concluded that the number of Internet users are increasing exponentially and so are the uses of Search engines. The users looking for information from World Wide Web (WWW) use search engine’s interface, where they provide search queries and the results thereof are displayed instantaneously on the screen in a ranked order. It generally happens that the entire results produced by search engines may not be relevant to the user. When the users do not get the desired information, they modify the search query again and again till they get the desired information or get tired. The situation becomes more cumbersome when the results produced by the search tools are outdated especially when a bad URL is reported.

Owing to the problems faced by users, a need arises to carry out some quantitative studies that can supplement in better understanding of the user behavior / requirement while using search engine and consequently helping to improve its working. Many surveys [99, 123-127] have already been carried out for Internet usage. These, however, do not emphasize all issues of user search trends and the problems faced in retrieving up-to-date information. In this work, a survey was conducted with a view to study the user search trends on WWW and an attempt was made to answer the following questions:

• How to refresh the web documents so that users get up-to-date information?
• How to improve the quality of repository of search engines such that irrelevant results while searching are reduced?
• What can be the remedy in improving the search trends?

A brief discussion on the survey carried out in this work is given in the following sections.

3.2 ANALYSIS OF USER SEARCH TRENDS ON INTERNET

3.2.1 SURVEY METHODOLOGY

During the survey, the user data was collected through a carefully designed questionnaire, uploaded online using Google docs wherein the questions were formulated keeping in view the user behavior towards putting the query, searching the information and getting the results. The target group consisted of people of wide variety from different walks of life with common traits
i.e. they search information on the Web. The following categories of users were selected for the survey:

- IT Professionals
- Academicians
- Non IT Professionals
- Bureaucrats

The questionnaire was sent to 6000 Internet users all over the world from above categories, whereas 560 responses in total from all categories were received. Based on the responses received following inferences were drawn:

3.2.2 INFERENCES DRAWN FROM THE RESULTS OF THE SURVEY CONDUCTED

Following are the observations made from the responses received and based on the responses inferences are drawn thereof:

**Q1:** On average how often do you use the Internet to search any information?
(a) Many times in a day  (b) Few times in a week  (c) Once in a week
(d) Once in a month  (e) Never use for searching

**Response:** As shown in Figure 3.8

**Observation:** It may be observed that 85.8% users agreed that they use Internet many times in a day while 14.1% use it few times in a week to search the required information and hence it can be said that almost 100% users nowadays use it to find the required information. It indicates the popularity of Internet and hence Search engines among users.

**Inference:** Internet is used frequently among users for searching information.

*Figure 3.8:* Use of the Internet
Q2: What do you use Internet mostly for?
   (a) Entertainment     (b) News     (c) Downloads     (d) Official/business
   (e) Shopping          (f) Communication  (g) Others
Response: As shown in Figure 3.9

Observation: Though Internet is used for variety of applications but few popular areas of applications are as shown in Figure 3.9. It may be concluded from the responses received, that mostly the users use Internet for official and communication purposes (65.85%).

Inference: Internet is used for variety of applications.

![Figure 3.9: Popular areas of applications of the Internet](image)

Q3: Which would you say best describes how you use search engines?
   (a) I always use the same search engine or web directory
   (b) I use multiple search engines interchangeably
   (c) I choose search engines according to information needed
Response: As shown in Figure 3.10.

Observation: Selection of search engines are not affected by the type of information required. As shown in Figure 3.10 approximately 79% users always use the same search engine. Only 12% users use multiple search tools whereas only 9% user’s use search tools according to the information required.

Inference: Selection of search engines are not according to the information needed.
Q4: Which of the following Navigation/Search Services do you primarily use?
(a) Google  (b) Alta Vista  (c) Yahoo  (d) Hotpot  (e) others
Response: As shown in Figure 3.11.
Observation: Approximately 98.21% internet users use Google whereas rest of the search engines are almost negligible whether it is AltaVista, Yahoo, Hotpot or any other search tools.
Inference: Google undoubtedly is the most popular search engine.

Q5. Average time taken by the search engine to produce result for a query is:
(a) Instantaneous  (b) Few minutes  (c) More than a few minutes  (d) Extremely long
Response: As shown in Figure 3.12.
**Observation:** Most of the Internet users get the search results instantaneously. Approximately 88% of the time, results are produced instantaneously and hence it proves the efficiency of Search engines in the current scenario. Cumulatively, 99% of time, results are produced within few minutes.

**Inference:** Search results are produced instantaneously.

---

**Q6.** How often do you find that the information returned by search engine is out of date?
(a) Never       (b) Seldom       (c) Sometimes       (d) Most of the times

**Response:** As shown in Figure 3.13.

**Observation:** Though users depend on Search engines for the required information but it did not always happen that the results produced are up-to-date with respect to its copy, existing at host server. Cumulatively, 48% users get out dated information occasionally as shown in Figure 3.13. Only 3% users claimed that they always get up-to-date results.

**Inference:** Results produced through search tools are not always up-to-date.
Q7: How many times do you encounter the broken links that do not work?
   (a) Never       (b) Seldom       (c) Sometimes       (d) Most of the times

Response: As shown in Figure 3.14.

Observation: Approximately 61% users feel that sometimes the results produced are not reachable i.e. they get broken links. Though there may be many reasons for getting broken links, one major reason is that the desired link no longer exists or gets changed. Cumulatively, 97% users face the broken links in some way or the other.

Inference: Results produced through search engines do no longer exist at host server end.

![Figure 3.14: Frequency of getting broken links](image)

Q8: How often do you feel the web document you visited gets changed?
   (a) Daily      (b) Weekly      (c) Monthly      (d) Yearly      (e) Don’t know

Response: As shown in Figure 3.15.

Observation: WWW is very much dynamic in nature and the same is reflected in the survey conducted. Approximately 42% users feel that the web documents get changed within a month and cumulatively, 61% user’s feel that the web documents get changed within a year.

Inference: Large fractions of web documents get changed.

![Figure 3.15: Frequency of web page getting changed](image)
Q9: How often do you return to already visited page?
(a) Never (b) Seldom (c) Sometimes (d) Most of the times

Response: As shown in Figure 3.16.

Observation: Approximately 16% Internet users revert back to already visited page most of the times whereas 62% users revert back sometimes during searching.

Inference: Users revisit already visited pages.

Figure 3.16: Frequency of reverting back to already visited web documents

Q10: Which of the following describes best what you are trying to do?
(a) I want to get to a specific website that I already have in mind.
(b) I want a good site on this topic, but I don't have a specific site in mind

Response: As shown in Figure 3.17.

Observation: 78% users responded that they want a good site but did not know any specific site. Only 22% users had specific sites in their mind. It shows that most of the users are not aware of any specific sites and therefore, blindly search the web.

Analysis: Most of the web users seek some good sites regarding their information but they are not clear about these sites.
Which of the following describes best what you are trying to do?

- 21% I want to get to a specific website that I already have in mind.
- 79% I want a good site on this topic, but I don't have a specific site in mind.

**Figure 3.17:** Users are not aware with any specific site

**Q11:** How often do you find the information you are looking for, on first page itself?

- (b) Never
- b) Seldom
- e) Sometimes
- (d) Most of the time

**Response:** As shown in Figure 3.18.

**Observation:** Only 5% users responded that they never get the desired information on the first result page, while 7% get it seldom on the first page. A figure of 26% users, sometimes get the information at the first result page, whereas, the rest 62% users, get the desired results most of the time on the first page itself.

**Inference:** *The users get the desired search results most of the times on the first result page.*

**Figure 3.18:** Users do get the desired results on first page

**Q12:** If you don’t find what you are looking for, at what point do you decide to change the search query?
(a) After looking first few entries at first page  (b) After looking first page results
(c) After looking 2-3 pages of result  (d) After looking more than 3 pages

**Response:** As shown in Figure 3.19.

**Observation:** Only 7% users go beyond the third page of results otherwise remaining 93% users get their search queries modified before exploring the results up-to the 3rd page. 70% users modify their search query after seeing the first page itself, out of which 26% explore only a few results from the first page before the search query gets modified, while 27% users go up to second page of results.

**Inference:** *The users of search engines do not search the result pages beyond the third page.*

![Pie chart showing search behavior](image)

**Figure 3.19:** Number of pages of search results seen before search queries get modified

**Q13:** On average after how many searches do you find the required information?

(a) 1-2 search  (b) 3-5 search  (c) 5 - 10 search  (d) More than 10 search

**Response:** As shown in Figure 3.20.

**Observation:** Approximately 83% users get the desired results within 1-5 searches, out of which 33% users agreed that they get the desired result within 1-2 searches itself. Only 17% users agreed that they get the desired results after 5 searches.

**Inference:** *Most of the users do get the desired result within first few searches.*
Q14: Do you feel that some of the search results returned by your search tool are not relevant to your query?
(a) Yes    (b) No

Response: As shown in Figure 3.21.

Observation: 37% users realized that they always get too many search results to browse. Cumulatively, 93% users agreed that some time or the other, they get undesired results to browse. It shows that users are not interested in thousands of search results returned by the search engine, rather they seek the desired information.

Inference: Most of the users feel that there are too many search results to browse.
Q15: What do you do when you do not get the desired results?

(a) Keep on searching  (b) Modify the search string (c) Switch to another search tool

Response: As shown in Figure 3.22.

Observation: 96% users prefer to modify their initial query when they do not get the desired search results, whereas only 4% switches to another search tool to get the information.

Inference: Most of the users, who do not get the desired results, prefer to modify the initial query instead of switching to any other tool or to keep searching.

![What do you do when you do not get the desired results?](image)

Figure 3.22: Users prefer to modify the search string rather than switching to another Search tool

Q16: What criteria do you use to write search string for searching information on the Web?

(a) Most frequently used words

(b) Most important words of the required information

(c) Specific topic or set of words

Response: As shown in Figure 3.23.

Observation: Cumulatively, 88% users responded that they either choose the most relevant words or the specific topic while writing the search query. Out of this, 54% stick to most relevant word, while 34 % prefer the specific topics.

Inference: Most of the users either chooses most important key words or specific topics related to information needed while providing the search string.
What criteria do you use to write search string for searching information on the web?

- 12% Most frequently used words
- 34% Most important words of the required information
- 54% Specific topic or set of words

Figure 3.23: The criteria used to write search string

Q17: Do you feel that there are too many search results to browse, returned by your search tool?

(a) Always  (b) Sometimes  (c) Seldom  (d) Never

Response: As shown in Figure 3.24.

Observation: 85% users realized that when they browse the search results, they come across irrelevant results, whereas remaining 15% feel that all results are relevant with respect to the search query fired.

Inference: Most of the users feel that some of the results returned by their search tool are not relevant to the query.

Do you feel that some of the search results returned by your search tool are not relevant to your query?

- 11% Yes
- 89% No

Figure 3.24: Some of the results returned are irrelevant
Q18: How often do you search images on the web?
   (a) Never          (b) Seldom          (c) Sometimes             (d) Most of the times

Response: As shown in Figure 3.25.

Observation: As far as image searching is concerned, only 10% users agreed that they use search tools, frequently for it and 7% users agreed that they never use search engine for it, whereas 83% agreed that they occasionally or rarely use search engines for searching images.

Inference: Image search are still not so popular as compared to text search.

![Figure 3.25: How frequently people use image search](image)

Q19: Are you aware of site maps?
   (a) Yes         (b) Yes, but never used       (c) No

Response: As shown in Figure 3.26.

Observation: 84% of internet users are aware with site map feature provided at the website but only 50% do use it, whereas the remaining 16% are not aware of this facility.

Inference: 50% Internet users are either not aware about the site map or are aware but never used it.

![Figure 3.26: Popularity of site maps](image)
Q20: How often do you use site map?
   (b) Often    c) Sometimes    d) Seldom    e) Never
Response: As shown in Figure 3.27.

Observation: Only 10% users use site maps frequently, whereas 26% use it occasionally and 23% use it rarely. 41% users had never used the site map feature. One of the reasons being that a very few number of websites do provide site map feature at their website.

Inference: *Exploring the required information through a site map is still unpopular.*

![Figure 3.27: How often site maps are used](image)

Q21: Do you know how to use the advanced search option?
   (a) Yes        (b) Yes, but never used      (c) No
Response: As shown in Figure 3.28.

Observation: Approximately 66% users do use advanced search features, facilitated by search engines, to refine their search results. Though, cumulatively 93% users are aware of this feature, out of which 27% are aware but had never used it as shown in Figure 3.28. Only 7% users responded that they are not aware of the advanced search feature.

Inference: Users also explore the advanced search features of search engines.

![Figure 3.28: Awareness about advanced search features](image)
3.2.3 SUMMARY

On the basis of the survey, the problems faced by the users in searching any information are discussed below:

- Users do not always get the relevant information on the first result page.
- Most of the users feel that there are too many search results to browse, returned by the search tool i.e. the problem of information overkill. It requires them to browse almost every web page in the returned list and sift the relevant information. In fact, the users are not interested in thousands of results, but want the most relevant information on the first or second page of search results.
- Sometimes users get outdated information i.e. the web documents at host server end get changed but the search engine’s database carries their older version and not the updated one.
- Sometimes users get broken links in their results i.e. either the links do not exist or get modified at host server but search engines still had those links in their database.
- Most of the users realize that some of the search results returned by the search tool are irrelevant.
- The users come across these irrelevant search results most of the time.

After analyzing the survey results and users search problems, the problems faced by the users in getting the relevant information on www are due to following factors:

- User Search Behavior
- Architecture of the current search tools

3.2.3.1 USER SEARCH BEHAVIOR

The search trends of the user that make the searching difficult are discussed below:

1. Most of the users still do not use advance search features to refine the results.
2. Users are still very much dependant on a single Search engine. They do not change the search tools even if they fail to get the relevant result in multiple searching.
3. Most of the users prefer to modify the initial query, if they are not getting the relevant information. However, they are not aware about the related keywords of the area in which they are seeking information, leading to irrelevant results.

3.2.3.2 ARCHITECTURE OF THE CURRENT SEARCH TOOLS

Web crawlers, download web documents from the Web and are responsible in building the database. However, they do not match their downloading frequency with the change frequency of web documents, owing to the following constraints:

1. Crawlers are not able to refresh the web documents on time.
2. Owing to sheer size and the dynamic nature of WWW, it is almost impossible to know the exact change frequency of all web documents.
3. The change detection techniques adopted by crawlers to find whether the web documents get changed or not are not so efficient to detect the changes. This is another factor that hampers the capability of crawlers in efficiently refreshing documents.
4. Also the crawlers working in parallel suffer from overlapping problem in the sense that multiple copies of the same web documents may be downloaded multiple times, leading to wastage of crawler’s time, network bandwidth and other resources such as storage at the Search engine side.

In the light of above discussion, it may be noted that there is a need to modify the existing architecture of web crawlers, with a view to improve the quality of downloaded web documents, within time constraints. In this work, a novel architecture for incremental parallel web crawler has been developed which not only downloads the more relevant documents at the earliest but also has a refresh module that detects whether a version of downloaded web document differs from its previous version or not, thereby helping in maintaining fresh copies of the documents at the Search engine side.