A Comparative Study of Google and Yahoo Web Resources on the Search term "Physics India"

Rasmita Mohanty K S Chudamani

1. Introduction

In recent years Internet has emerged as the most important and powerful medium for storage and retrieval of information. It works 24×7 hrs and connects every nook and corner of the globe, thus being treated as the biggest open library of the world. In today's world information transfer through web plays a significant role in the utilization of its resources, thus understanding of their structure and formats is essential.

There is a tremendous growth in the number and variety of information resources available on the Internet and it made a great impact on information. Over the past 10 years due to the buzz of open access and open access movement online information becomes an important source for scholarly scientific literature and also more number of sources as well as the results of scientific research is now being available on web. Thus identifying their scholarly characteristics and new potential users has become important. Moreover, the study of how scholars use and disseminate information on the web through formal and informal channels has created new opportunities to access online or web based science communication paradigm changes. The present study is based on what are the major resources and types of resources and new web types and their formats available on the web for research communication .For this a sample search of 100 hits were taken under the search term "Physics India" in two of our major search engines, i.e. Google and yahoo.

2. Web Resource: The Concept

Online or electronic information is becoming a major factor in information activities not only in developed countries but also in developing countries. Information architecture as an emerging discipline encompasses the design and maintenance of electronic spaces (E-Spaces) with an emphasis on access and usability. Due to the rising trends of electronic documents the use of cyberspace becomes popular. The concept of "web resource" is being used interchangeably synonymous with online resource, digital resource and e-resources. But in simple connotations web resource can be regarded as the resource, document or information available on the Internet or World Wide Web.

The concept of "resource is primitive in the web architecture, and is used in the definition of its fundamental elements. The term web resource was first introduced to refer to targets of uniform resource locators (URLs) but its definition has been further extended to include the referent of any Uniform Resource Identifier (URI) (www.wikipedia.org). According to Vishnu Kant Sukla a web

resource can be defined as a resource which is present on Internet in the electronic form or we may say that, the resources located remotely and can be accessed through interactive communication with the help of computer and communication channel.

A web resource or webpage is an unit of information often called a document that is available over the world wide web .Web resources are created using HTML, which defines the contents of a webpage such as images, text, hypertext links, video and audio files etc. Web resources are sent and received through HTTP, a method used to transfer hypertext files across the Internet .Information contained in the web resources are provided in the form of hypermedia pages , which combines graphics and text and also have the added feature that users can follow the links provided to other documents located virtually anywhere on the web.

3. Web resource: Basic Features

The following are the basic features of a web resource:

- Web resources are accessed and browsed using HTTP protocol and files are exchanged using FTP
- Created using HTML
- Interactive in nature
- Posses international reach /wider accessibility
- Speed of communication
- Unlimited capabilities
- Reduced cost
- Search ability and
- Linking

4. Search Engine

In simple words a search engine is a software that searches through a database of web pages or web resources for a piece of information, keywords, concepts etc.

To define the concepts more descriptively we can say that "search engine is a computer program that searches for documents containing words or phrases of interest to users .The search engine itself is a virtually powerful workstation-class machine that searches a database of information collected from the Internet. Primarily software program called robots or spiders that crawl through all the files on the Internet and download them into a searchable database .These works as indexes to the literature available on the network (Hussain & Kumar, 2006).

There are a number of search engines available on the web. Most of the search engines provide website reviews and homepage services in addition to keyword searches .A number of studies have been carried out in the past which compares the search and retrieval features of various search

engines .But in this present study two most popular search engines have been studied in terms of its available web resources with reference to Physics-India in Google and Yahoo

4.1 Google: An Overview

Google was created in the winter of 1998 by graduate students at Standard University and was officially launched in the fall of 1999.This is a straightforward engine that does not support advanced search syntax making it very easy to use and retrieves pages ranked on the basis of number of sites linking to them and how often they are visited, indicating their popularity (ibid). It claims that 97% of the users find what they are looking for.

Features

Google includes the following most important features:

- Cached page archives
- Result clustered by indention
- Result displayed option, from 10-100

"Google Search" Supports:

- Implied Boolean (+)sign, (-) sign
- Double quotes ("") for phrases
- Stop words.

Other Search Options Available with Google:

- "I' m Feeling Lucky"(goes directly to top ranked site in query)
- "Google scout" (bring up list of related sites)
- "Uncle Sam" (Searches govt. and Milsites)
- "Search within results" option
- Field searching with 'link' only

(http://www.google.com) (Hussain & Kumar, 2006)

4.2 Yahoo: An Overview

Yahoo is a subject Directory and also a commercial portal compiled by human. It is oldest as well as largest directory on the web launched in mid 1994. This is one of the most frequently accessed tools, and although most people consider it as a search engine , it is basically classified as a directory(Chowdhry, 2004).

Yahoo allows the user to put a search query, its strength lies in the categories and each that can lead a user step-by-step to the desired subject category. At present it has 26 categories, about 315+ Sub Categories; Sub-sub-categories can be estimated as more than 700 excluding the BEST ANSWERS Category. (http://answers.yahoo.com/question/index?qid=20071215131412AAsf3ZB)

Structure

- Yahoo is hierarchically organized with subject catalogue or directory of the web which is browsable and searchable.
- Links to various services are accomplished in two ways such as
  - by user's submissions and
  - Through robots that retrieve new links from known pages.
- Yahoo indexes web pages, UseNet and e-mail address

Features

- Topic and region specific "yahoos!"
- Automatic truncation
- No case sensitivity and stop words
- The syntax that yahoo follows for searching is fairly standard among all search engines

Search Option

Users can browse Yahoo! Simply by clicking on the various categories listed on each page, or can search Yahoo! By entering a word into the search box that appears on every page in the directory. Again one can combine the two strategies and can "browse and then search" or "search and then browse."

The following are the various search facilities available in yahoo.

"Main page" supports:

- search in yahoo's subject categories
- implied Boolean(+) and (-) signs
- double quotes(" ") for phrases i.e. phrase search
- Truncation: use of * e.g. physic*, denotes suffix or right truncation.
- Field specific search: use of (t :) and URL respectively.

Advanced search (labeled 'search options') supports:

- All features of "main page" search and Boolean type searching.
- Yahoo subject categories.
- "UseNet news groups" searches
- date range searches, from 1 day to 4 days
- result displayed from 20 to 100

Other search options

- Yahoo! News
- User may combine any of the query syntax as long as the syntax is combined in the proper order, which is +, -, t:, "", and *.If Yahoo does not find any matching entries, pertaining to a query, in its main database, the query will automatically be transferred to the Inktomi database,

a search engine that automatically 'crawls' the text of the entire web. Inktomi database contains results for literally millions of individual web pages.

Yahoo thus looks for information in:

- Yahoo! Categories
- Websites listed in yahoo
- WebPages indexed by Inktomi.

(Chowdhry; p404)

5. Web Resources on Physics India

While carrying out this study, the prime goal was to know the various kinds of resources available on the web on the broad subject Physics in various Indian domains as well as the most commonly available formats and the characteristics of the resources with their frequency of occurrence retrieved through two major search engines taking into account 100 hits among each.

Thus while making out the study the characteristics of the web resources has been classified into the following 16 major categories as:

- Journals
- Journal articles
- Books
- Book chapters
- E-prints
- Conference papers
- Discussion forums
- Databases
- Pointer pages (links to websites)
- Web directories
- Research news
- Associations
- Videos
- News clips
- Personal news
- Conferences

Based upon the search query "Physics India" the domains has been classified into 8 and studied in relation to the above kinds of web resources as:

- .ac
- .com
- .org
- .res
- .ernet
- .gov

- .net
- .edu

The file formats were classified into two broad groups based upon the availability of majority of resources on it .These are mentioned below:

- PDF
- Html.

6. Web Resources: A Comparative Study

In the present study six excel sheets were prepared to compare the total number of hits in relation to the types of web resources vs available domains and available formats among the first hundred hits.

6.1 Web resources of Google

Table 6.1 gives a sketch of web resources on "Physics India" retrieved through the Google search out of 100 links.

Result: The analysis of the data available with the table 6.1 shows that most of the web resources under the search term "Physics India" retrieves the pointer pages (links to websites on the same subject) and achieves 67% among all the other kinds of resources. Secondly relates to the journal articles by 26% of retrieved output. And the lowest percentage of search results deals with research news, news clips, databases and conference papers .Table 6.1.1Shows the graphical representation of the retrieved output through Google.

Table. 6.1 web resources vs frequency of their occurrence per search

Kinds of Web Resources	Number of searches										Total
	1	2	3	4	5	6	7	8	9	10	
Journals	1	1	2	2	2	1	1	2	2	1	15
Journal articles	2	5	5	4	5	1	1	1	1	1	26
Books											0
Book chapters											0
Eprints	1										2
Conference papers	1										1
Conferences											4
Discussion forums	1										6
Databases	1	1	1								1
Pointer pages	1										67
web directories	1	1	2	2							2
Research news					1						1
Associations	42	1	2	1	1	4	5	2	4	5	1
videos				1		1					5
news clips	1										1
personal news	1										2
	1	1	3								
						1					
		1	1								

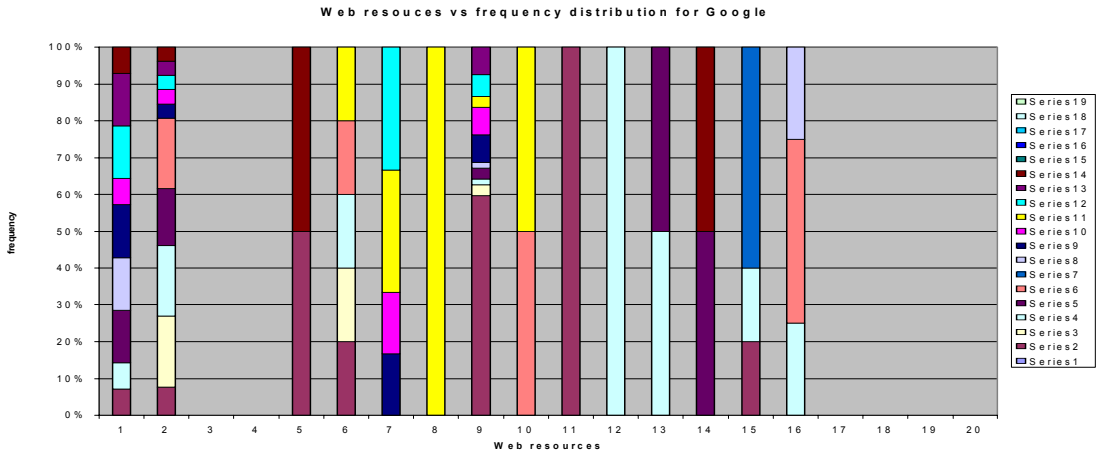

Fig. 6.1.1 web resources vs frequency of their occurrence per search

6.2 Web Resources of Yahoo

Table-6.2 shows the ratio of web resources on "Physics India" retrieved through the Yahoo Search. The figure 6.2.1 provides the graphical representation of the frequency of occurrence of various kinds of web resources.

Result:

The analysis and interpretation of the data available in the table reflects that most of the retrieved results provides pointer pages with a percentage of 27% and second comes the web directories as 18% and it provides lowest percentage of retrieval rates of Journal articles.

Table 6.2 web resources vs frequency of their occurrence per search

Kinds of Web Resources	Number of searches										Total
	1	2	3	4	5	6	7	8	9	10	
Journals		2		1	2	1		1			7
Journal articles					1						1
Books					2					1	3
Book Chapters	1	1									2
Eprints	1		1	1						1	4
Conference papers											0
Conferences	1	1			2						4
Discussion forums			1	1				1	1	1	5
Databases		1	1				1	1	1		5
Pointer pages	5	3	2		2	1	3	3	3	5	27
web directories	1		1	2	2	1		3	3	5	18
Research news											0
Associations			1							1	2
videos											0
news clips	2		1	1	2		2	2	2	1	13
personal news											

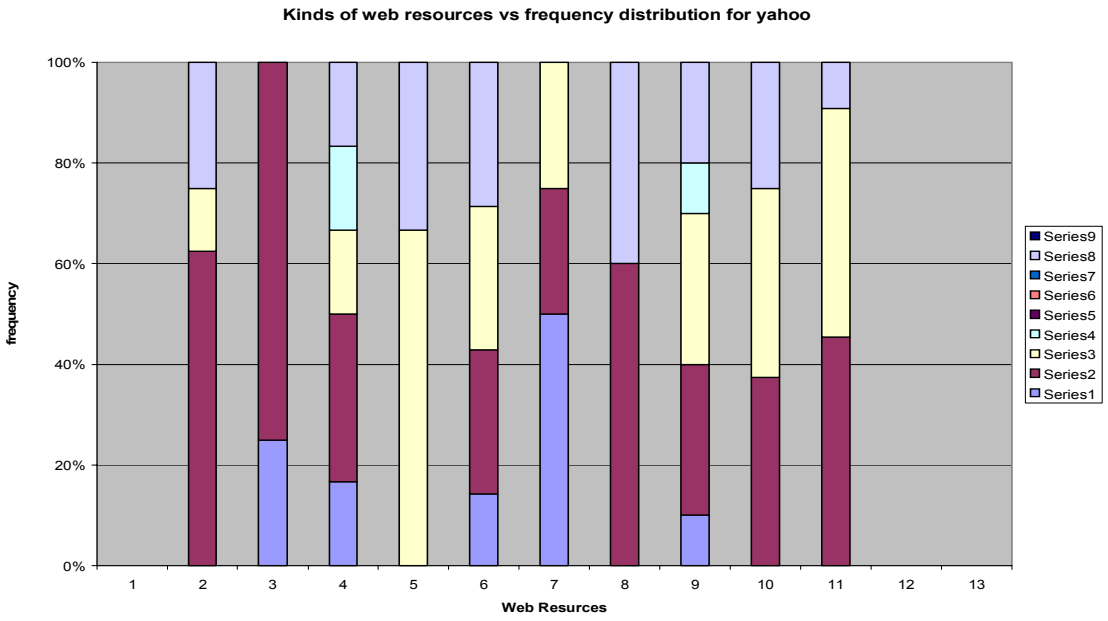

Fig. 6.2.1 web resources vs frequency of their occurrence per search

6.3 Domains of Google

Here, while making a search under the search term "Physics India" through the Google it is being found that majority of the resources on Physics is available in under mentioned eight main domains .Table 6.3 delineates the major domains and the frequency of occurrence of the resources on those and Fig 6.3.1 provides the graphical representation of the frequency of occurrences.

Table. 6.3 Domains vs frequency of their occurrence per search

Main domains	Serial Number of searches										
	1	2	3	4	5	6	7	8	9	10	Total
.com	2	3	2	1	1	3	6	5	4	5	32
.ac	5	6		9	9	1	1	1	1	2	35
.edu	1		1								3
.net	1										3
.res	1		1								7
.org	1										15
.gov	3		1				1	1	1		4
.ernet	1	3		2	1	4		1	2	1	6
	1		2				1				
			1	2				1	1	1	

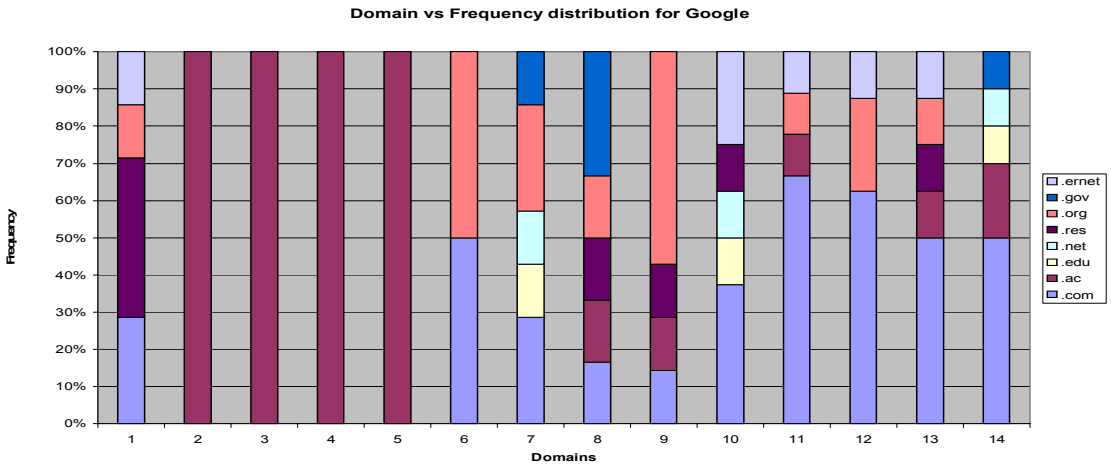

Fig 6.3.1 Domains vs frequency of their occurrence per search

Result: We found that 35% of sources on physics were from academic domains and 32% of sources were from commercial domains. But the lowest percentage of resources were from sub group of the academic domains ending in .edu or net e.g. (.edu, .net) as well as from government domains

6.4 Classification of Domains on Yahoo

Similar to the above classification of domains and the frequency of occurrence of the sources, the Table 6.4 shows the major domains and the frequency of occurrence of the resources on those and Fig 6.4.1 provides the graphical representation of the frequency of occurrences.

Table.6.4 domains vs frequency of their occurrence per search

Main domains	Serial Number of searches										
	1	2	3	4	5	6	7	8	9	10	Total
.com	3	6	5	4	8	2	6	4	7	4	49
.ac		2				1	1	2			6
.edu		1	1	1	1	2	1	1	2	3	13
.net											
.res	1	1	1	1							4
.org	2		2	4	1	4	2	3	1	3	22
.gov								1			1
.ernet	3	1									4

Result: The data from the above table reflects that most of the resources on the physics are available in commercial domains and secondly on organizational domains of India. And very lowest percentage indicates to the government sites.

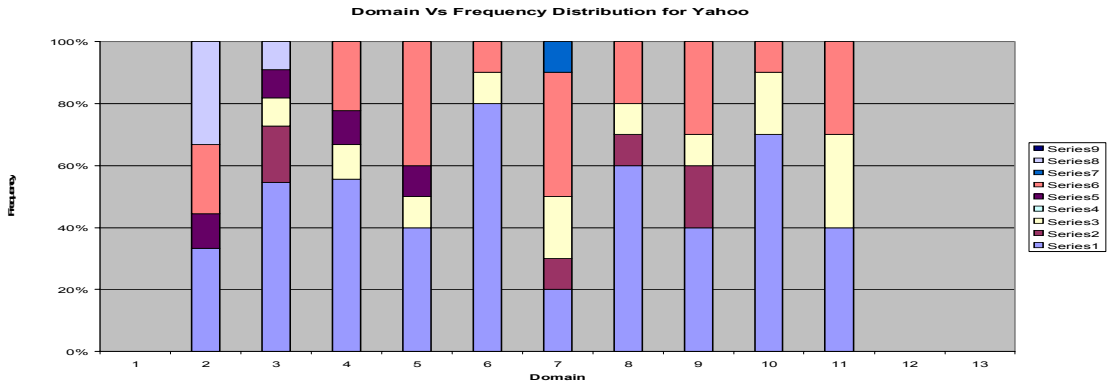

Fig. 6.4.1 domains vs their frequency (Yahoo)

6.5 Classification of the file formats : Google

While carrying out the study we found that there are two main file formats on which almost all of the resources on Physics are available on the web retrieved through Google and Yahoo. Thus Table 6.5 indicates the file formats and the frequency of the resources on that and Fig 6.5.1 shows the graphical representation of the frequency distribution.

Table.6.5 File formats vs frequency of their occurrence per search

File formats	Serial Number of searches										
	1	2	3	4	5	6	7	8	9	10	Total
PDF	1	3	6	2	6	8	2	6	2	2	38
HTML	1	3	7	5	7	1	2	1	1	2	30

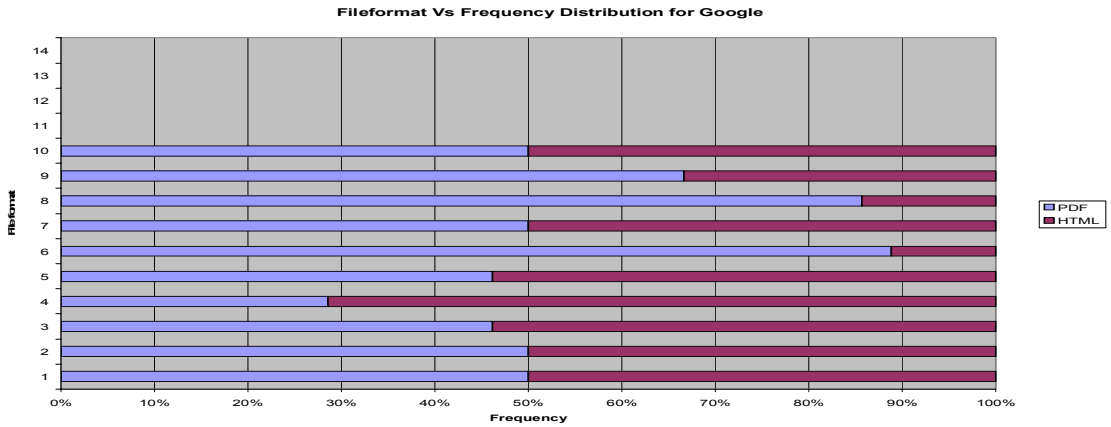

Fig 6.5.1 file formats vs their frequency (Google)

Result: Above data indicates that maximum resources on physics retrieved through the Google are available in PDF (Portable Document Format).

6.6 File Formats: Yahoo

Table 6.6 indicates the file formats and the frequency of the resources on that and Fig 6.6.1 shows the graphical representation of the frequency distribution.

Table 6.6 File formats vs frequency of their occurrence per search

File formats	Serial Number of searches										
	1	2	3	4	5	6	7	8	9	10	Total
PDF		2	1	1		1	1	1			7
HTML	1	4			1			2	1		9

Result: Here from the above data it is clear that most of the web resources on Physics India retrieved through Yahoo search is on HTML format.

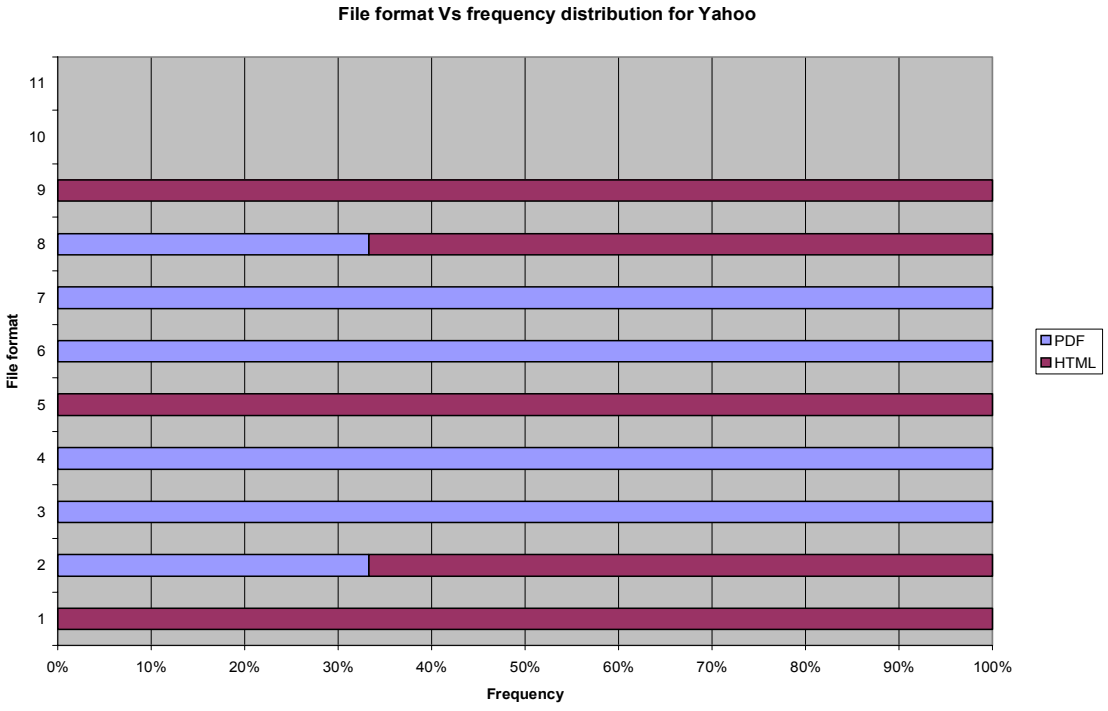

Fig.6.6.1 File formats vs frequency of their occurrence per search

7. Conclusion

The analysis of the results of the above calculations of the data suggests that the web contains a wide range of resources on Physics from India which provides only links to other web pages in the same subject. But the only difference is that the percentage of links to other pages (Pointer pages) is high in Google search than that of Yahoo search.

While wide range of resources retrieved through Google search is available in academic domains (.ac), most of the resources retrieved through yahoo search used to be available on Commercial domains.

And most of the resources retrieved through Google search are available in PDF formats but most of the resources retrieved through Yahoo are available in HTML formats.

It is also been found while searching through both of the search engines on the same query "Physics India" the same site reappears on several pages, which reduces the relevancy of the retrieved output. Overall the search result of the Google retrieves more number of resources on the Physics while Yahoo retrieves less number of sources in comparison to Google. As well as it has been found that the results of Google haves more relevancy on the context than that of Yahoo results.

Finally, it has been experienced that to obtain the most useful results from Googles and yahoo's URL statistics, it is necessary to develop algorithms and or deploy human labor to avoid the reappearing of the same sites or sources and then to separate out the different kinds of sites. Though the same suggestion has also been provided by Kousha and Thelwall on their study "How Science Cited on the Web? A classification of Google Unique Web Citations. This has great implications for use of web resources.

References

1. Hussain, Akthar and KUMAR, Krishna. Search Engines: An Overview. ILA Bulletin. 2006, 42(3), p. 21-26.
2. Kousha, Kayvan and THELWAL, Mike. How Science Cited on the Web? A classification of Google Unique Web Citations. Journal of the American society for Information Science and Technology. 2007, 58(11), p.1631-1644.
3. Sukla, Vishnu Kant. Inclination of Scientists towards e-information in the libraries of CSIR Institutions of Luck now. Herald of Library Science .Jan-Apr 2005, 44(1-2), p.53-60.
4. Saravanan, T and PONNUDURAI, R. Reports on the potential aspects of research in Astronomy in G7 Countries: A Bibliometric analysis. IASLIC Bulletine.2006, 51(3), p.169-177.
5. Mounissamy, P and KALIAMMAl, A. Promoting effective use of Electronic resources using library websites by IITs and NITs: A Comparative Study. IASLIC Bulletin. 2006, 51(4), p.213-220.

6. Koovakkai, Dineshan and NOOR HANA, K V. Electronic information use among the faculty. Library Herald. December 2006, 44(4), p313-320.
7. Sing, Rajesh. Performance of World Wide Web Search Engines: A Comparative Study. New Delhi: Delhi Library Association, 2006. p. (328-338).
8. Chowdhary, G G. Introduction to Modern Information Retrieval. Great Britain: Facet Publishing, 2004. p (395-404).
9. http://answers.yahoo.com (accessed on 10/1/2008)
10. http://www.google.com (accessed on 25/10/2007)
11. http://www.yahoo.com (accessed on 12/10/2007
12. http://www.wikipedia.org (accessed on 15/10/2007)

About Authors

Ms. **Rasmita Mohanty,** Trainee, JRD Tata memorial Library, Indian Institute of Science, Bangalore. E-mail: rasmita06@gmail.com.

Dr. K S Chudamani, Deputy Librarian, JRD Tata memorial Library, Indian Institute of Science, Bangalore.