Search Engines Analysis for Search Result Based On Phrase (Group of Words) As Query Topic

Nripendra Dwivedi, Ajay Tripathi, Ashwani Varshney

Abstract*: A huge number of search engines are accessible at the web to fulfill the users' need based on their query topic. There is a challenge for web users to take decision about the selection of appropriate search engine. Search engines return the quality web documents in proper order decided by their specific ranking algorithm in the searching process of any query string. Web search users are very much interested about the quality outcome of search result. We are considering query topic in form of group of words having some specific meaning. In order to find suitable search engine, a study of four distinct selected search engines (Yahoo, Google, Ask and Bing), decided by the survey of internet users on many distinct queries is done. The parameter for analysis taken into consideration is: "number of web links in search result against given query topic framed with group of words like phrase". Analysis of variance is used to derive the result in the research process. In this research paper, it is also analyzed that considered search engines are significantly different or not.

***Index Terms*: ANOVA, F Test, Ranking Algorithm, WWW,**

I. INTRODUCTION

The World Wide Web (WWW) has become a huge information storage of millions of inter linked WebPages. It fulfills the need of the people around the globe. Searching the required result from huge storage efficiently is big challenge of web users. Efficient search engine can solve this problem by searching appropriate data.

Search engine is one of the most widely used applications of internet family [1].It reflects the result based on the search query

As per research result of Marchionini[2,3,8] "end users want to achieve their goals with a minimum of cognitive load and a maximum of enjoyment", It shows that many internet users see the result within first web page in nthe searching process.

By Gwizdka Jacek and Chignell Mark [4], title "Towards Information Retrieval Measures for Evaluation of Web Search Engines," Department of Mechanical and Industrial Engineering , University of Toronto, CANADA 1999, three search engines AltaVista, HotBot, infoseek were used in this research for comparison. Different evaluation measures Precision, Presentation - Ranking, User Effort - Search length, Coverage, hit & hit ratio were used for analysis. AltaVista was found to have better performance in terms of precision.

By Liu Bing [5], "User Personal Evaluation of Search Engines-Google, Bing and Blekko," as per "Technical Report, Department of Computer Science, University of Illinois at Chicago, May 8, 2011", the evaluation shown that in terms of user satisfaction, Google was found the best, Bing was close behind.

In January 2006, a research done by IProspect [6] about the behavior of web users, it is found that "Key among the finding relating to the current search engine community is that 62% search engine user's click on search result within the first page"

The relative study (analysis) of Yahoo, Google, Ask and Bing is shown in this research paper based on outcome about number of relevant link with exact query string in first web page of all search result. ANOVA tool is used in the study of these selected popular search engines.

II. QUALITY PARAMETER

The quality parameter which is considered for evaluation of search engines is mentioned below:

Number of web links in search result against given query phrase framed with group of words

This quality parameter shows the total number of relevant link appearance with exact query string in first return web page result. With the help of many links, user may be benefitted by finding many results for any specific query phrase.

III. METHODOLOGY

For comparative analysis of these selected search engines, 150 meaningful topics (query strings) in form of phrase (group of words) are taken from different subject areas. The query strings (phrases) were selected such that they represent some normal (specific) meaning. All these topics reflect some standard concept. These topics are written in the coupons separately. After that, all coupons are scuffled and put up in container. Out of 150 keywords from different IT area, only 30 keywords are picked randomly from container using lottery system. The various topics are shown in Table I.

We have also selected four popular search engines on the basis of survey. Survey is done by taking opinion of 140 internet users from different

Revised Manuscript Received on June 05, 2019

Nripendra Dwivedi,Professor &Associate Dean- Research, Department of Computer Science, IMS, Ghaziabad,Dr Abdul Kalama Technical University Lucknow

Ajay Tripathi is Associate Prof. at Jaipuria Institute Of Management, NCR,Delhi

Ashwani Varshney is Associate Prof. at Jaipuria Institute Of Management, NCR,Delhi

places. We picked top four search engines as popular search engines on the basis of survey result. Hence, four popular search engines decided for analysis are Google, Yahoo, Bing and Ask.

Searching process was started with the help of four considered search engines on each and every topic. Total number of relevant links of first web page result out of all returned web page results through search engines was maintained corresponding to every topic in tabular form.

On this quality parameter (Number of relevant links in the result for given query topic framed with group of words like phrase).We would like to do analysis to check the equivalency of all considered search engines. For this analytical process, ANOVA and F test concepts are used. Mean of number of links through Google, Yahoo, Bing and Ask is shown in Table I. Grande mean of these means with respect to these four search engines is also displayed in Table I. Calculation of between column variance is shown in Table II and within column variance is shown in Table III. "Between column variance" and "Within column variance" are also calculated using appropriate formulas, shown in the Table II and Table III correspondingly.

With the help of Table I, the value of "between column variance" is found and Table II gives the value of "within column variance".

Now, the value of F is derived as per given formula below

F=Variance of web links ("between column")/Variance of web link ("within column") = 4.474 / 3.758165= 1.191

It is also shown in the ANOVA table in summarized form.

We got the degree of freedom from ANOVA table is 3 and 116 for between column sample and within column sample respectively. The F (tabular value) of (3,116) is 2.74 for 5% significance level. The value of F (calculated value) is found with the help of Table 4 as 1.19. Thus, the calculated value (F=1.19) is less than that of tabular value as (F=2.74). On the basis of this fact, we can say that assumption of null hypothesis i.e. ($\pi_{1=}$ $\pi_{2=}$ $\pi_{3=}\pi_4$) is true, ($\pi_{1,}$ $\pi_{2,}$ $\pi_{3,}$ π4 are means of samples corresponding to considered four search engines). Consequently, it is derived that search engines (Yahoo, Google, Bing and Ask) in terms of returning search result do not differ significantly

IV. KEY FINDING &RESULTS

On the basis of analysis in the research, it is found that all selected popular search engines (Yahoo, Google, Bing and Ask) do not differ majorly in terms of total number of links matching with exact query string-phrase, returned in first web result. In terms of result of web links, they are treated as equivalent.

V. CONCLUSION

The World Wide Web (WWW) is global source for retrieval of variety of informational entity. Huge number of web documents is added with new content in the World Wide Web at every hour. This is the reason; efficient search engine is required for searching. Every search engine is returning the quality result with the help of their confidential intelligent ranking algorithms according their query string- phrase very effectively.

A bulk of the web consumer sees the first web page outcome. Hence, by putting eminent results at opening web page based on highest rank against group of words as query topic become very useful to web users. As per the derived result in this study, Search engines viz Yahoo, Google, Bing, Ask are treated as equivalent in terms of result of web links.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table 1: No OF LINKs FOUND CORRESPONING TO RESPECTIVE SEARCH ENGINE

Topic	Number of relevant web link(reference) in the web search result with exact matching of query string through Google	Number of relevant web link(reference) in the web search result with exact matching of query string through Yahoo	Number of relevant web link(reference) in the web search result with exact matching of query string through Bing	Number of relevant web link(reference) in the web search result with exact matching of query string through Ask
Graph Terminology & Representations,	**2**	**1**	**1**	**2**
Graphs & Multi-graphs,	**1**	**1**	**1**	**0**
Minimum Cost Spanning Trees.	**3**	**2**	**2**	**4**
Connected Component and Spanning Tree	**0**	**1**	**1**	**1**
Sequential Representations of Graphs,	**0**	**1**	**1**	**0**
Two Way Merge Sort,	**4**	**5**	**5**	**3**
Practical consideration for Internal Sorting.	**2**	**0**	**0**	**0**
Sorting on Different Keys,	**1**	**1**	**1**	**0**
Hash Table Implementation.	**4**	**2**	**2**	**3**
comparison and analysis,	**1**	**3**	**3**	**1**
Overflow and Underflow,	**4**	**7**	**7**	**2**
Binary Search Tree	**9**	**9**	**9**	**9**
Binary tree representation,	**3**	**5**	**5**	**3**
Two Way Merge Sort,	**4**	**5**	**5**	**3**
Collision Resolution Strategies	**1**	**3**	**3**	**1**
Array and Linked Representation of Binary trees,	**0**	**0**	**0**	**0**
B+ Tree index files,	**3**	**4**	**4**	**3**
Queries and Sequential organizations,	**0**	**1**	**1**	**0**
B Tree index files	**1**	**2**	**2**	**1**
Searching and Hashing:	**1**	**1**	**1**	**1**
Traversing Threaded Binary trees,	**0**	**0**	**0**	**0**
random and linked organization	**0**	**0**	**0**	**0**
Data Structure operations,	**2**	**2**	**2**	**2**

Algorithm Complexity and Time-Space trade-off.	**0**	**0**	**0**	**0**
Linear and Multidimensional Arrays,	**1**	**1**	**1**	**1**
Representations of stack,	**2**	**4**	**4**	**0**
Primary, secondary and hash indexing	**0**	**0**	**0**	**0**
AVL Trees,	**9**	**9**	**9**	**9**
Polynomial representation and addition	**2**	**2**	**2**	**0**
Garbage Collection and Compaction.	**1**	**1**	**1**	**0**
Mean	2.033333	2.433333	2.433333	1.633333
Grade mean=(30/120)*2+(30/120)*2.43333+(30/120)*2.4333+(30/120)*1.6333 = 2.1249				

Table 2 Population variance

No of link found by searching through Google (Method1) (z1)	$(z1-mean(z1))^2$	No of link found by searching through Yahoo (Method2) (z2)	$(z2-mean(z2))^2$	No of link found by searching through Bing (Method3) (z3)	$(z3-mean(z3))^2$	No of link found by searching through Ask (Method4) (z4)	$(z4-mean(z4))^2$
2	0.00111089	2	0.18774889	2	0.1877489	2	0.13446889
1	1.06770889	1	2.05434889	1	2.0543489	1	0.40106889
3	0.93450889	3	0.32114889	3	0.3211489	3	1.86786889
0	4.13430889	0	5.92094889	0	5.9209489	0	2.66766889
0	4.13430889	0	5.92094889	0	5.9209489	0	2.66766889
4	3.86790889	4	2.45454889	4	2.4545489	4	5.60126889
2	0.0111089	2	0.18774889	2	0.1877489	2	0.13446889
1	1.06770889	1	2.05434889	1	2.0543489	1	0.40106889
4	3.86790889	4	2.45454889	4	2.4545489	4	5.60126889
1	1.06770889	1	2.05434889	1	2.0543489	1	0.40106889
4	3.86790889	4	2.45454889	4	2.4545489	4	5.60126889
9	48.5349089	9	43.121549	9	43.121549	9	54.2682689
3	0.93450889	3	0.32114889	3	0.3211489	3	5.60126889
4	3.86790889	4	2.45454889	4	2.4545489	4	5.60126889

1	1.06770889	1	2.05434889	1	2.0543489	1	0.40106889
0	4.13430889	0	5.92094889	0	5.9209489	0	2.66766889
3	0.93450889	3	0.32114889	3	0.3211489	3	5.60126889
0	4.13430889	0	5.92094889	0	5.9209489	0	2.66766889
1	1.06770889	1	2.05434889	1	2.0543489	1	0.40106889
1	1.06770889	1	2.05434889	1	2.0543489	1	0.40106889
0	4.13430889	0	5.92094889	0	5.9209489	0	2.66766889
0	4.13430889	0	5.92094889	0	5.9209489	0	2.66766889
2	0.0111089	2	0.18774889	2	0.1877489	2	0.13446889
0	4.13430889	0	5.92094889	0	5.9209489	0	2.66766889
1	1.06770889	1	2.05434889	1	2.0543489	1	0.40106889
2	0.0111089	2	0.18774889	2	0.1877489	2	0.13446889
mean=2.03333	103.2586332	mean=2.4333	110.4812711	mean=2.433	110.481271	mean=1.6333	111.762791
	variance of sample $s1^2=(z1-mean(z1))^2/(29)=$ 3.56065517		variance of sample $s2^2=(z2-mean(z2))^2/(29)=$ 3.809655		variance of sample $s3^2=(z3-mean(z3))^2/(29)=$ 3.809655		variance of sample $s4^2=(z4-mean(z4))^2/(29)=$ 3.853897
Within column variance=$\sum((n-1)/(n-c))*s_i^2=((29)/(120-4))*3.56066+ (29/116)*3.809+ (29/116)*3.809+(29/116)*3.854=3.758165$ where, n-c=Degree of freedom							

Table 3 : Between column variance using four Search Engines

Number of search query strings in the sample i.e. size (N)	a=mean of number of links corresponding to query topics through every search engine	Grand mean of means corresponding to four search engines	u^2 =a-(Grand mean)^2	N*u^2
30	2	2.1249	.0156 (Google)	0.468
30	2.4333	2.1249	.3084^2=0.09511 Yahoo	2.853
30	2.4333	2.1249	.3084^2=0.09511 Bing	2.853
30	1.6333	2.1249	0.492 Ask	7.248
				Total=13.422
between column variance=$\sum$ n*u^2/t-1=13.422/(4-1) =4.474 where, t-1=Degree of freedom				

Table 4: ANOVA Table

Variation Type	Summation of square	Degree of freedom	Mean square	The value of F (computed result)
Between(among) sample	13.422	(4-1)=3	4.474	4.474/3.759= 1.19021
Within sample	435.979	(120-4)=116	3.75844	

REFERENCES

1. Courtois, Martin P., Baer, William M., and Stark, Marcella. Cool tools for searching the Web: A performance evaluation. Online, 19(6), 14-32
2. Marchionini, G. 1992. "Interfaces for End-User Information Seeking." Journal of the American Society for Information Science,43(2):156-163.
3. Krishna Bharat, Monika R. Henzinger , "Improved Algorithms for Topic Distillation in a Hyperlinked Environment" 21st ACM SIGIR Conference, 1998
4. Gwizdka Jacek, Chignell Mark, "Towards Information Retrieval Measures for Evaluation of Web Search Engines," Department of Mechanical and Industrial Engineering, University of Toronto, CANADA, 1999
5. Liu Bing, "User Personal Evaluation of Search Engines-Google, Bing and Blekko," Technical Report, Department of Computer Science, University of Illinois at Chicago, May 8, 2011
6. IProspect, Jupiter, "iProspect Search Engine User Behavior Study," A report of Research, January 2006
7. Angelo Chianese, Fiammetta Marulli, Francesco Piccialli, Paolo Benedusi, Jai E. Jung,"An associative engines based approach supporting collaborative analytics in the Internet of cultural things", Elsevier journal -Future Generation Computer Systems, vol. 66, Pages 187-198, January 2017
8. N. Dwivedi, L. Joshi, & N. Gupta. Statistical analysis of search engines (Google, Yahoo and Altavista) for their search result. International Journal of Computer Theory and Engineering. 2013, 5(2), 298-301
9. Soper, D.S) critical f-value calculator [software].available from http://www.denielsoper.com/staticalc, 2017
10. Wilfred Amaldoss, Preyas S. Desai, Woochoel Shin,"Keyword Search Advertising and First-Page Bid Estimates: A Strategic Analysis" Journal of the Management Science,2015

AUTHORS PROFILE

Dr Nripendra Dwivedi is working as Professor (Comp.Sc) with responsibility Associate Dean-IT & Associate Dean-Research in IMS, Ghaziabad, Dr Abdul Kalama Technical University Lucknow. He has done**Ph.D**(Comp.Sc area from Govt university ,University of Rajasthan, Jaipur), M.Tech (Comp.Sc-79%) (**Gold Medalist**), M.Phil(Comp.Sc-79%), MCA and B.Sc.He is also **Gold Medalist**(95%) in NPTEL- RLanguage from **IIT Kanpur** by scoring 95%**.** He has published around **50 research papers** in referred National/International Journals/Conferences viz. IEEE Xplore, Journal of Science and Technology, International journal of computer theory & engineering ,Scopus Indexed Journals etc and has presented papers at various National/International Conferences/Seminars. He is having around 16 years of Experience in academia as well as Industry. He is also life time member of reputed professional organization IETE (Institute of Electronics and telecommunication Engineers) He has been honored as session chairperson in different reputed International conferences (including International conference organized by IMT, Ghaziabad etc). He has worked as Head-MCA, Chairperson-Computer Centre and Web Master in different reputed organizations. He also contributed by taking responsibility as Head Examiner (Many times) for evaluation of answer sheets of B.Tech/MCA of entire Mahamaya technical University &UP Technical University.He has been achieved various awards like NCC (Gold Medalist), National Scholarship holder etc. Various workshops and seminars on advanced technology are conducted by him. He is writing two books. It would be shortly published. His area of interest is Data Structure, .Net frame work, Advance Database and Automata Theory.

Dr Ajay Tripathi is Associate Prof. at Jaipuria Institute of Management, NCR,Delhi

Dr. Ashwani Varshney is Associate Prof. at Jaipuria Institute of Management, NCR,Delhi