Statistical Analysis of Search Engines (Google, Yahoo and Altavista) for Their Search Result

Nripendra Dwivedi, Lata Joshi, and Neeraj Gupta

***Abstract*—The web search engines are used to extract query specific information from this massive pool of WWW. A large number of different search engines are available to the user to satisfy their needs. Every search engine uses its own specific algorithm to rank the list of web pages returned by the search engine for the users query, so that the most relevant page appears first in the list. Users think which search engine should be selected for searching corresponding to any query topic for efficient search. For decision making on the basis of search result users want to know, whether they are significantly different or not. In this regard, a study of three popular different search engines (i.e. Google, Yahoo and Altavista) on twenty-four different query topics in terms of the quality of most relevant web page is done. The three parameters of the quality of the most relevant web page taken into consideration are: 'in depth coverage', 'ideation and clarity', and 'number of links of related articles to query topic'. For each query topic, the first web page returned by these three search engines were evaluated and graded by the experts of that specific area. Based on these grades, a table is maintained showing the relative performance of selected three search engines in terms of three parameters of quality of first returned web page and thus comparing the ranking algorithms used by them. We analyzed the result using analysis of variance and F test. By this analysis we found grade for first return page result corresponding to each and every topic through all these three search engines (Google, Yahoo and Altavista) does not vary significantly. Thus they are equivalent. So users can treat these three poular search engines as equivalent search engine for the first return page result on the basis of these three parameters.**

***Index Terms*—Search engines, world wide web, google, yahoo.**

I. Introduction

In the past decade the world has witnessed the explosion of World Wide Web from an information repository of few million of hyper linked documents into a massive world wide "organism" that serves informational, transactional and communicational needs of people all over the globe. Though a latecomer in the Internet family, it has rapidly gained popularity and became the second most widely used application of the Internet [1]. Search Engines are specially designed for informational retrieval, which extracts the information from WWW as per users query. As argued by Marchionini, [2], [3] "end users want to achieve their goals with a minimum of cognitive load and a maximum of enjoyment", correspondingly, in the context of web searches it is observed that a maximum of the search engine users tends to click on a result within the first page of the search results. In fact, a survey done by IProspect and Jupiter research on the behavior of search engine users in January 2006 shows that 62% of the search engine users click on a search result within the first page of results [4]. Since, generally search engines returns a very large list of documents for the users query, the list is ranked in accordance with the importance and relevance to the users query. Thus, for the users query, the ranked list of results is displayed by search engine, with few results per page.

The general concept used by most of the search engines to find quality web pages and rank the list that is used by Page Rank algorithm [5], [6] which assumes that if web page A has a hyper link to web page B then the author of web page A thinks that web page B contains valuable information. This opinion of A becomes more important. This means that ranking of a web page is high if many highly ranked web pages points to it.

This paper gives the focus for comparative study (analysis) of these search engines (Google, Yahoo and Altavista) in term of quality of first return page result. ANOVA(analysis of variance)[7] and F test[8] is used for comparative statistical analysis of these search engines. On the basis of this analysis, research would show about equivalence of these search engines.

The rest of the paper is organized as follows: Section II describes quality parameter on which these search engines were evaluated to check their efficiency. Section III describes the methodology followed in carrying out the research. Section IV describes the key finding and result and Section V concludes the paper.

II. Quality Parameter

The quality parameters on which the search engines were evaluated as follows

A. In Depth Coverage

The parameter "In Depth Coverage" refers to the level of depth up to which the description of sub topics is given.The description of every topic is done in terms of its building blocks or sub topics. The description of subtopics is also done in the same manner and the process continues until the basic building blocks are reached. While describing a topic, a quality web page includes a hyper link to its sub topics and clicking these hyper links takes the user to web pages where these sub topics are described. This is helpful and required if the end user is not aware with the technical concepts and various building blocks used while describing a topic.

Manuscript received September 6, 2012; revised November 20, 2012.

Nripendra Dwivedi is with the University of Rajasthan, Jaipur, India (e-mail: ohmdwivedi@hotmail.com).

Lata Joshi is with the Department of Statistics, University of Rajasthan Jaipur, India (e-mail: lj_statistics@yahoo.com).

Neeraj Gupta is with the Department of Comp. Sc(ITS) G. B. Technical University, India (e-mail: neerajgupta.842001@gmail.com).

B. Ideation and Clarity

This quality parameter refers to the clarity with which the concept is formulated and presented. The completeness of sentences, proper use of grammar, unambiguous sentences, the flow of idea, continuity, and division of the text into proper paragraphs is an issue here.

C. Number of Links of Related Articles to Query Topic

A good web page about a topic not only describes the topic but also provides links to related documents about which the author of the web page thinks that they are also highly valuable and can be referred by the end user. This quality parameter refers to the number of such related articles provided. This is helpful to the user in the way that using this user can see many different articles by different authors.

III. Methodology

To do the comparative study of these search engines, randomly four specialized areas were selected. From each of these four specialized areas, randomly three different sub specializations were selected and from each of these, two different topics were selected arbitrarily. The various areas and topics are summarized in Table I.

TABLE I: The Various Specialized Areas, Sub Specialized Areas and Topics Selected for Their Grade of Two Search Engines.

SPECIALIZED AREAS	SUB SPECIALIZED AREAS	TOPICS
Computer science	Data structure	Bubble sort
		B tree
	Data base management system	3NF
		Super Key
	Software Engineering	Unit Testing
		COCOMO Model
Physics	Mechanics	Force
		Moment of Inertia
	Modern Physics	$E=mc^2$
		Nuclear fusion
	Optics	Reflection
		Critical angle
Chemistry	Biochemistry	Mitrocandria
		RNA structure
	Organic Chemistry	Phenol
		Picric acid
	Physical Chemistry	PH value
		Titration
Mathematics	Abstract Algebra	Semi Group
		Sub ring
	Graph Theory	Shortest path Algorithm
		Binary tree
	Trigonometry	Area of triangle
		Identity

The topics selected were such that they represent standard concepts and have standard meaning. These topics do not represent a broader area and cannot be further bifurcated into further broad sub areas.

Search was initiated on each of the topics on the three selected search engines. The first document returned was saved from the ranked list of documents. For each of these documents, experts of the corresponding areas graded them on the scale of 0 to 10, with 0 being the least and 10 being the best, on each of the three quality parameters, separately. Corresponding to every topic and each search engine, the three quality grades were averaged to get the average quality. All average grades corresponding to respective topics are shown in Table II.

TABLE II: Grades of Corresponding Topics for Each search Engine

Area	Topics	Grade of search result corresponding to Google Search Engine (in 1 to 10 scale)	Grade of search result corresponding to Yahoo Search Engine (in 1 to 10 scale))	Grade of search result corresponding to Altavista Search Engine (in 1 to 10 scale))
Computer Sc	Bubble sort	7.7	7.7	7.7
	B tree	8.7	8	8.7
	3NF	6.7	6.7	6.7
	Super Key	5.7	5.7	5.7
	Unit Testing	6.3	6.3	6.3
	COCOMO Model	7.7	7.7	7.7
Physics	Force	8	8	8
	Moment of Inertia	7.3	7.3	7.3
	$E=mc^2$	8	8	8
	Nuclear fusion	7.7	7.7	7.7
	reflection	3.7	6.7	6.7
	Critical angle	7.3	7.3	7.3
Chemistry	mitochondria	8.3	8.3	8.3
	RNA structure	2.7	8.7	2.7
	Phenol	8.3	8.3	8.3
	Picric acid	7.7	7.7	7.7
	PH value	8.3	8.3	8.3
	Titration	7	7	7
Mathematics	semigroup	7.7	6.3	7.7
	Sub ring	6	0.3	6
	Shortest path algorithm	8.3	7.6	8.3
	Binary tree	6	6	0.3
	Area of triangle	6.3	6.3	6.3
	Identity	0.3	0.3	1.7
		Mean(x1)= 161.7/24=6.74	Mean(x2)=162.2/24=6.758	Mean(x3)= 160.4/24=6.683
GRAND MEAN(y)=(24/72)*6.74 +(24/72)*6.76 + (24/72)*6.683= 6.727				

On these parameters (In Depth Coverage, Ideation and Clarity, Number of links of related articles to query topic) constraint, for analysis point of view, we want to check whether these search engines are significantly different or not. For that we apply Analysis of variance test and F test. Table II shows mean of grade through method 1(Google search engine), method2(Yahoo search engine) and

method3(Altavista search engine). Table III also shows grand mean of grade by these three methods. By using appropriate formula we calculate population variance(between column variance) and population variance(within column variance) shown in Table III and Table IV respectively. Table III and Table IV also show appropriate formula for required calculation.

TABLE III: SHOWS POPULATION VARIANCE BETWEEN COLUMN VARIANCE THROUGH BOTH SEARCH ENGINE.

n(size of each sample ie number of search topics taken as query topic)	**ā=Mean of grade of all topics through every search engine**	**Grand mean of mean grade through all these three search engines**	**y^2=(ā –Grand mean)^2**	**n*y^2**
24(by google)	6.74(by google)	6.727	$(-0.013)^2$= 0.000169 (of google)	24*0.000169= 0.004056 (of google)
24(by Yahoo)	6.76 (by Yahoo)	6.727	$(0.033)^2$=0 .001089 (of Yahoo)	24*0.001089= 0.026136 (of Yahoo)
24(by Altavista)	6.683 (by Altavista)	6.727	$(-0.044)^2$ =0.001936 (of ALTAVIS TA)	24*0.001936= 0.046464 (of Altavista)
				$\sum n_j y^2 = \sum n_j$(ā–Grand mean)^2=0.07 6656
Between column variance= $\sum n_j$(ā –Grand mean)^2/(k-1)= 0.076656/(3-1)= 0.038328 Where k=number of sample nj =size of jth sample				

Null Hypothesis**:-** We assume sample means corresponding to these three search engine is identical. $\mu1=\mu2=\mu3$ where, $\mu1$= sample mean corresponding to Google search engine. $\mu2$=sample mean corresponding to Yahoo search engine. $\mu3$= sample mean corresponding to Altavista search engine

F-test: - An F-test is any statistical test in which the test shows the presumed null hypothesis is true or false. The hypothesis states that multiple normally distributed populations, all having the same standard deviation, are equal. This is perhaps the most well-known of hypotheses tested by means of an F-test, and problem in the analysis of variance (ANOVA).

F hypothesis test=Analysis of variance compares these two estimates of the population variance by computing their ratio called F as follows

F=Estimate of population variance based on variance among sample means (between column variance)/ Estimate of the population variance based on the variances within the samples (within column variance).

We can find the value of between column variance and within column variance using Table III and Table IV respectively. It is also shown in ANOVA table (Table V)

TABLE IV: SHOWS POPULATION VARIANCE- WITHIN COLUMN VARIANCE THROUGH BOTH SEARCH ENGINES.

Grade for search result against Google Search Engine (Method1) (x1)	**(x1-mean(x1))^2**	**Grade for search result against Yahoo Search Engine) (Method-2) (x2)**	**(x2-mean(x2))^2**	**Grade for search result against Altavista Search Engine) (Method-3) (x3)**	**(x3-mean(x3))^2**
7.7	(7.7-6.74)^2= 0.9216	7.7	(7.7-6.76)^2= 0.8836	7.7	(7.7-6.68)^2 = 1.0404
8.7	3.8416	8	1.5376	7.3	0.3844
6.7	0.0016	6.7	0.0036	6.7	0.0004
5.7	1.0816	5.7	1.1236	5.7	0.9604
6.3	0.1936	6.3	0.2116	6.3	0.1444
7.7	0.9216	7.7	0.8836	7.7	1.0404
8	1.5876	8	0.09	8	1.7424
7.3	0.3136	7.3	0.2916	7.3	0.3844
8	1.5876	8	0.09	8	1.7424
7.7	0.9216	7.7	0.8836	7.7	1.0404
3.7	9.2416	6.7	0.0036	1	32.2624
7.3	0.3136	7.3	0.2916	1	32.2624
8.3	2.4336	8.3	2.3716	1	32.2624
2.7	16.3216	8.7	3.7636	2.7	15.8404
8.3	2.4336	8.3	2.3716	0.3	40.7044
7.7	0.9216	7.7	0.8836	7.7	1.0404
8.3	2.4336	8.3	2.3716	8.3	2.6244
7	0.0676	7	0.0576	6	0.4624
7.7	0.9216	6.3	0.2116	7.7	1.0404
6	0.5476	0.3	41.7316	0.3	40.7044
8.3	2.4336	7.6	0.7056	8.3	2.6244
6	0.5476	6	0.5776	6	0.4624
6.3	0.1936	6.3	0.2116	6.3	0.1444
0.3	43.4056	0.3	41.7316	0.3	40.7044
Mean(x1)= 161.7/24=6.74	$\sum$(x1-mean(x1))^2 =93.5884	Mean(x2)= 162.2/24=6.76	$\sum$(x2-mean(x2))^2 =103.2832	Mean(x3)= 160.4/24=6.683	$\sum$(x3-mean(x3))$^{2=}$ 250.5792
	Sample variance = $s1^2$= ($\sum$(x1-mean(x1))^2)/(24 -1)= 93.5884/23= 4.0691		Sample variance = $s2^2$ = ($\sum$(x2-mean(x2))^2)/(24 -1)= 103.2832/23= 4.4906		Sample variance= $S3^2$= ($\sum$(x3-mean(x3))^2)/(24 -1)= 250.5792/23= 10.8947
Population variance(within column variance)= $\sigma^2=\sum((n_j-1)/(n_t-k))sj^2$ =((24-1)/(72-3))* 4.0691+(23/69)* 4.4906 + (23/69)* 10.8947=6.4848 Where n_j=size of jth sample, n_t=Total sample size, k= number of sample, sj= sample variance through jth search engine					

Putting these values in above formula, we find,

F=Between column variance / Within column variance =0.038328 / 6.4848= 0.00591

ANOVA table shows degree of freedom corresponding to

within sample and between samples are 2 and 69 respectively.

TABLE V: ANOVA TABLE SHOWS-VARIANCE RATIO.

Source of variation	Sum of Square	Degree of freedom	Mean Square	Variance Ratio(F)
Between Samples	0.076656	3-1=2	(0.076656/2)=0.038328	0.038328/ 6.4848= 0.00591
Within Samples	447.4512	72-3= 69	(447.4512/69)=6.4848	

The tabular value of F for (2, 69) at 5 % level of significance is 3.15 Since the computed value of F=0.00591 is less than the tabular value of F=3.15, therefore we accept our null hypothesis i.e. our null hypothesis assumption ($\mu1=\mu2=\mu3$) is true, ($\mu1$, $\mu2$, $\mu3$ are sample means corresponding to these three search engines).So we can say search result through these search engines (Google, Yahoo and Altavista) will not differ significantly.

IV. KEY FINDINGS AND RESULTS

By Analysis of variance(ANOVA) and F test, it is concluded that these search engines can be considered equivalent in view of these three parameters (In Depth Coverage, Ideation and Clarity, Number of links of related articles to query topic) of first return result (document) for any query topic. Hence these three search engines do not vary significantly.

V. CONCLUSION

The web has become "the place" for accessing any type of information. There are billions of web pages and, everyday, new content is produced. Therefore, the use of search engines is becoming a primary Internet activity, and search engines have developed increasingly clever ranking algorithms in order to constantly improve their efficiency. There are many search engines available and their actual ranking algorithms are not made available to the rest of the world. Because a majority of the end users see the first return result. So by providing quality result at highest rank search engines become effective and popular.

By consideration of "In Depth Coverage, Ideation and Clarity, Number of links of related articles to query topic" if we evaluate Google search engine , Yahoo and Altavista search engine we found these three search engines do not differ significantly. (In Depth Coverage, Ideation and Clarity, Number of links of related articles to query topic)The web search engines are different, in various aspects, from the well established other search tools. Therefore, they require a different evaluation methodology, and we have made an attempt with three popular search engines and twenty-four different sample queries. In this research, we have evaluated the three search engines on the basis of these parameters of search result corresponding to sample queries.

In the future, we plan to apply the proposed methodology to a wider scope with the hope that our research findings will truly enable web users to select a search engine on basis of (In Depth Coverage, Ideation and Clarity, Number of links of related articles to query topic) parameter, and help web search engine developers design even better ones for the Internet community.

REFERENCES

[1] P. C. Martin, M. B. William, and S. Marcella, *Cool Tools for Searching the Web: A Performance Evaluation*. Online, vol. 19, no. 6, pp. 14-32.

[2] G. Marchionini, "Interfaces for end-user information seeking," *Journal of the American Society for Information Science*, vol. 43, no. 2, pp. 156-163, 1992.

[3] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," Presented at 21st ACM SIGIR Conference, 1998.

[4] iProspect Search Engine User Behavior Study. (January 2006) A Report by iProspect and Jupiter Research. [Online]. Available: http://www.iprospect.com

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The page rank citation ranking: Bringing order to the web," *Stanford Digital Libraries Working Paper*, 1998

[6] T. H. Haveliwala, "Topic sensitive page rank," *in Proceedings of the Eleventh International World Wide Web Confrernce*, 2002.

[7] One Way ANOVA - University of Wisconsin - Stevens Point. [Online]. Available: http://www.uwsp.edu/psych/stat/12/anova-1w.htm

[8] An Outsource Specializing in Functionality Testing of Web Sites. Multimedia, CD-ROMs, and Internet Applications. [Online]. Available: http://www.chem.utoronto.ca/coursenotes/analsci/StatsTutorial/ftest.h

Mr. Nripendra Dwivedi is from India and date of birth is 2nd Aprl 1975.He has done B.Sc,MCA(ComputerApplication),M.Phil(Comp.Sc), M.Tech(AlCCS) (Comp.Sc) (secured Highest rank) from different reputed organizations and pursuing Ph.D from university of Rajasthan, Jaipur, India He has published more than 25 research papers in referred National/International Journals/Conferences and has presented paper at various National/International Conferences/Seminars.

Dr. Lata Joshi is from Rajasthan, India. She has done B.Sc,M.Sc(2nd Rank in University of Rajasthan Jaipur),M.Phil,Ph.D. She is having more than 30 years of Experience as academician. She has published many research papers in referred National/International Journals/Conferences.

Mr. Neeraj Gupta is with the Computer. Sc. Department (ITS), G. B. Technical University, India .