Determining the Most Significant Factors in Classifying a Web Site – Users Perspective

B. B. Jayasingh[1] and Nayani Sateesh[2]

[1] Professor, IT Dept., CVR College of Engineering, Ibrahimpatan, RR Dist-501510.
Email: bbjayasingh9@rediffmail.com

[2] Asst. Professor, IT Dept., CVR College of Engineering, Ibrahimpatan, RR Dist-501510.
Email: nayanisateesh@gmail.com

***Abstract*- Due to the wide range of connectivity, communication over Internet becomes a single click. In this way, websites play a vital role to reach and attain the users. Hence, we must design the website considering the features the users look and the features that are most significant to them. In this work, we try to understand the user behavior and the features that are most important to them in order to classify a website as an effective website. The user behavior depends on the usage and accessing the websites in a given time. We propose to apply a statistical approach called Factor Analysis to identify the most significant features the user considered for classification. We plan to conduct a survey on various types of users by selecting the respondents using purposive random sampling. In this Survey, We conduct a pilot survey to understand the features the users generally look at.**

***Index Terms*- Purposive Random Sampling, Factor Analysis. Principal Component Analysis (PCA), Sampling methods, Website features.**

I. INTRODUCTION

The world is becoming a global village due to the connectivity provided by the Internet. In this global village environment, people are connecting over the Internet to connect, share and exchange the information or services across the world. Internet will provide the interface to connect each and everyone via a web site. Hence web site is playing a vital role to reach and attain the users. Such users may be general purpose users or business oriented.

We are trying to identify the features that the users will look when they are using the website and the features that are most significant to them in classifying the web site as an effective website using Factor Analysis [1, 7].

Understanding the user requirements is more critical in successful development of an application. Reaching the customers and attain the customers is crucial in the business community when the Internet is becoming the way of communication mode. In this way of communication websites plays a vital role to attain the customers and retain them for longer times and increase the customer base. In this work, we are training to understand the user behavior and hence to understand what features are more important to them when they are using or accessing the websites and what feature make them to classify a website an effective website[7].

We are following statistical approach, in which we are conducting a survey on various types of users by selecting the respondents using purposive random sampling. In this Survey, We are conducting a pilot survey first to understand what features users generally will look at. Based on the users' feedback, we are trying to identify the most significant features they considered for classification using Factor Analysis. Factor analysis will help us to identify the most significant factors for classification.

II. BACKGROUND

As part of this work, we are trying to identify what are the features the users will look when they are using the website and what features are most significant to them in classifying the web site as an effective website using Factor Analysis [1,7].

A. Web Site Features

Consistency in Design: A website should be designed in such a way that all pages and forms should be designed uniformly in appearance and functionality which improves the user experience. When navigating

DOI:10.32377/cvrjst0514

to other pages, users should feel comfort with reference to appearance of the pages and page elements.

User-friendly: The website should be designed in such a way that a normal user will be able to navigate and find the information likes to view. Navigation should be simple and properly designed the linkages between the pages.

Perfect Content: Content is the prime factor upon which the users are engaged on to a web site for a longer time. Information or the content published on the web site should be complete, correct, concise, accurate and updated. It should be free from grammatical errors. Background colors and highlighted portions should be paid more attention. Content should be properly formatted. Too much complicated formatting and highlighting should be avoided.

Fast Loading: Try to avoid the inclusion of heavy elements like Graphics, animations, videos etc. even though they add more creativity to site but may cause the delay in loading when requested by the users. Create the site with the simple design since users may not have patience to wait for longer time.

Search Engine Friendly: Web site should be designed in such a way that the elements in web site should help in good raking of the website on different search engines. It helps to reach the more users on the web through the search engines.

Compatible on Different Browser: Web site should be tested on different versions of the web browsers and platforms in order to verify whether the web site is working properly or not since some of elements on web page may not work properly as intended.

Functionality: Poorly constructed website may frustrate the users. Avoid the page errors. Navigation should be properly designed. Links should properly make among the web pages. Need to check the functionality of all the pages and elements to improve the quality of the site.

B. Sampling Methods

Sampling methods are categorized into either probability sampling or non-probability sampling [4, 8]. Every sample point of the population has non-zero probability of being selected in probability sampling where as in non-probability sampling, sample point has the random probability of being selected. Probability sampling methods include random sampling, stratified sampling systematic sampling. Non- probability sampling includes quota sampling, convenience sampling, snowball sampling and purposive or judgment sampling.

Random sampling: Each observation has the known and equal probability of being selected. This sampling method helps to avoid the biased nature of the data.

Systematic sampling: From the population every n^{th} element is selected as the sample point and constitutes the sample. Advantage in this sampling is simplicity.

Stratified sampling : Stratum is the subset of the population which shares the common features. Based on the feature stratums are identified. Random sampling is used to select the subjects from each stratum. This sampling method used frequently when stratums have low incidence relative to other.

Convenience sampling : is used in exploratory research In this method sample is selected based on convenience. This method is used in explorative research studies in getting an inexpensive approximation of the truth.

Quota sampling :it is equivalent to stratified sample where the stratum is filled by random sampling. Stratums and their proportions are identified as they represent the total population.

Snowball sampling : It is used when the sample under study has the rare characteristic. It may be extremely difficult to locate respondents in these situations. It relies on referrals from initial subjects to generate additional subjects.

Judgment sampling / Purposive Sampling: sample is selected based on judgment. This sample is considered to be true representation of the entire population.

III. PURPOSIVE RANDOM SAMPLING

The process of identifying a population of interest and developing a systematic way of selecting cases that is not based on advanced knowledge of how the outcomes would appear. It unable to identify and differentiate the need of the various related groups. It generates the sample where the included groups are selected based on specific characteristics considered to be important. With such a sample, group differences can be compared and contrasted and a range of experiences can be summarized.

We conducted a pilot survey to understand the significance factures in users' perspective in classifying the websites [5]. Based on the features mentioned by the respondents we conducted the original survey using purposive random sampling with sample size 57. The following table shows the template of the metadata that is being collected for the analysis.

DOI:10.32377/cvrjst0514

TABLE I
VARIABLE LIST USED IN ANALYSIS

Variable Name	Data Type	Description
Rand_ID	Numeric	User ID
Gender	String	Sex of the Respondent
Age	Numeric	Age of the Respondent
email	String	Respondent Contact Email Id
Time_Spent	Numeric	No. of Minutes Spent on Internet A Day
Most_Visit_Web	String	Most Frequently Used Web Site
Var_Feature1_Rank	Numeric	Rank for Influence Factor (Consistency in Design)
Var_Feature2_Rank	Numeric	Rank for Influence Factor (User-friendly)
Var_Feature3_Rank	Numeric	Rank for Influence Factor (Perfect Content)
Variable Name	Data Type	Description
Var_Feature4_Rank	Numeric	Rank for Influence Factor (Fast Loading)
Var_Feature5_Rank	Numeric	Rank for Influence Factor (Search Engine Friendly)
Var_Feature6_Rank	Numeric	Rank for Influence Factor (Compatible on Different Browser)
Var_Feature7_Rank	Numeric	Rank for Influence Factor (Functionality)

Data is collected from a sample (internet users) using purposive random sampling based on their opinion on the features which influence them in classifying the web sites as effective web sites [5]. Data given in the table is percentages. N represents Total Number of respondents i.e 57. Out of 57 respondents, there are 31 Female and 26 male respondents.

TABLE II
USER DATA

Features/Factors	Gender (N=57)		TotaL (N=57)
	Female (N=31)	Male (N=26)	
Consistency in Design	3	8	5
User-friendly	13	27	19
Perfect Content	26	8	18
Fast Loading	29	19	25
Search Engine Friendly	10	15	12
Compatible on Different Browser	13	12	12
Functionality	6	12	9

The following graph shows the influence of the feature with reference to gender comparison. The numbers shown in the graph are in percentages.

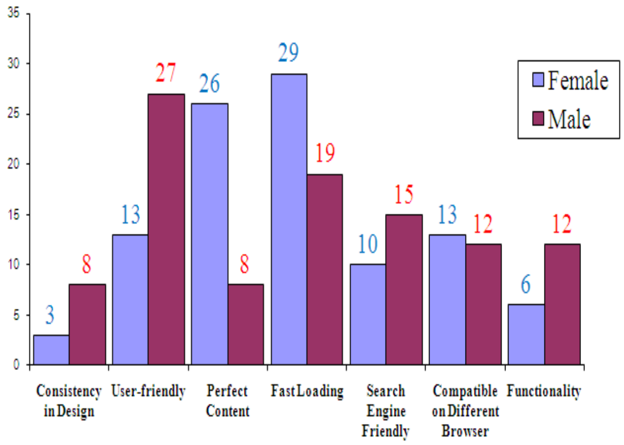

Figure 1: Feature Vs Gender Influence

The following graph shows how features are influencing irrespective of the gender. We observed that fast loading, perfect content and user friendly features are influencing the classification of web sites as effective (The numbers shown in the graph are in percentages.)

DOI:10.32377/cvrjst0514

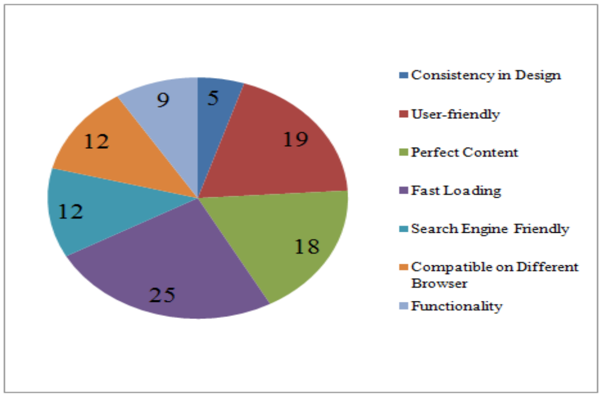

Figure 2. Feature Selection

The following test data shows that there is a significance difference in usage of internet between the Male and Female users.

Here we have used mean and standard deviation in order to calculate the mean difference between the male and female users to test whether there is any significance difference between the usages of the internet. Standard error mean is estimated by the sample estimate of population standard deviation.

TABLE III
DESCRIPTIVE STATISTICS ON INTERNET USAGE.

Group Statistics					
	Gender	N	Mean	Std. Deviation (S)	Std. Error Mean
No. of Minutes Spent on Internet A Day	Female (Y_1)	31	55.71	34.597	6.214
	Male (Y_2)	26	75.19	44.102	8.649

t-test is used to test whether there is a significance difference between two sample means. Here is the t-test to calculate the equality of means i.e between the internet usage of female vs male respondent . We used the t-statistic formula as follows.

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$$

Where $y_1$and y_2 represents response of female and male respondent's respectively. S_1 and S_2 represent the standard deviation of the female and male respondent's response respectively. N1, N2 represents total number of female and male respondents respectively.

TABLE IV
TEST OF SIGNIFICANCE FOR USAGE OF INTERNET.

Levene's Test for Equality of Variances			
		F	Sig
No. of Minutes Spent on Internet A Day	Equal variances assumed	3.738	0.058

TABLE V
t-TEST FOR SIGNIFICANCE OF EQUALITY OF MEANS

t-test for Equality of Means						
		Sig.			95% Confidence Interval of the Difference	
t	df	(2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
-1.869	55	.067	-19.483	10.426	-40.376	1.411

Here we are not considering the F statistic, significance of Levine's Test. t-represents for t-statistics for calculating the mean difference. At 55 df (degrees of freedom) with 95% confidence level we came to know that there is a significance difference in the usage of the internet between the Male & Female respondents. We found that the significance value (0.067) is greater than critical value (0.05) which rejects the null hypothesis that means there is a significant difference between the mean usages of the internet.

IV. FACTOR ANALYSIS

Factor analysis is a statistical procedure used to identify a small number of factors that can be used to represent relationships among sets of interrelated variables. Factor analysis is used in many areas, and is of particular value in psychology, sociology, market research and education[3]. For example, COMPUTER USE BY TEACHERS is a broad construct that can have a number of FACTORS [1] (use for testing, use for research, use for presentation development, etc.).

DOI:10.32377/cvrjst0514

Multiple linear regression model:

$$x_1 = \lambda_{11} f_1 + \cdots + \lambda_{1k} f_k + u_1$$
$$x_2 = \lambda_{21} f_1 + \cdots + \lambda_{2k} f_k + u_2$$

. .
.= .
. .

$$x_p = \lambda_{p1} f_1 + \cdots + \lambda_{pk} f_k + u_p$$

where
x = (x_1…..x_p)’ are the observed variables (random)
f = (f_1…..f_2)’ are the common factors (random)
u = (u_1…..u_2)’ are called specific factors (random)
λ_{ij} are called factor loadings (constants)

BASIC ASSUMPTION: underlying dimensions – or factors – can be used to explain complex events or trends.

Objective: It is to identify otherwise not-directly-observable factors on the basis of a set of observable variables.

FOUR STEPS:

1. Compute a correlation matrix for all variables.
2. Determine the number of factors necessary to represent the data and the method of
Calculating them (factor extraction)
3. Transform the factors to make them interpretable (rotation)
4. Compute scores for each factor.

Factor analysis usually proceeds in two stages [1]. In the first, one set of loadings is calculated which yields theoretical variances and covariances that fit the observed ones as closely as possible according to a certain criterion. These loadings, however, may not agree with the prior expectations, or may not lend themselves to a reasonable interpretation. Thus, in the second stage, the first loadings are “rotated" in an effort to arrive at another set of loadings that fit equally well the observed variances and covariances, but are more consistent with prior expectations or more easily interpreted.

A method widely used for determining a first set of loadings is the principal component method [2]. This method seeks values of the loadings that bring the estimate of the total communality as close as possible to the total of the observed variances.

V. RESULTS

The following tables show the Eigenvalues and the factor loading which are calculated in factor analysis based on Principal Component Analysis (PCA) [6].

TABLE VI
TOTAL VARIANCE EXPLAINED

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	1.71	24.427	24.427	1.71	24.427
2	1.425	20.352	44.779	1.43	20.352
3	1.201	17.159	61.938	1.2	17.159
4	1.001	14.294	76.231	1	14.294
5	0.925	13.219	89.45		
6	0.738	10.55	100		
7	-1.00 E-13	-1.02 E-13	100		

TABLE VII
TOTAL VARIANCE EXPLAINED

Component	Extraction Sums of Squared Loadings	Rotation Sums of Squared Loadings		
	Cumulative %	Total	% of Variance	Cumulative %
1	24.427	1.46	20.861	20.861
2	44.779	1.316	18.805	39.667
3	61.938	1.286	18.377	58.044
4	76.231	1.273	18.188	76.231

The following tables show the significance factor loading with respect to each component. We need to consider the maximum factor loading for the corresponding component which in turn influences the classification.

DOI:10.32377/cvrjst0514

TABLE VIII
FACTOR LOADING OF THE COMPONENTS.

Component Matrix				
Factors	Component			
	1	2	3	4
Compatible on Different Browser	-0.73	-0.31	0.339	
Fast Loading	0.58	0.253	-0.53	-0.3
User-friendly	-0.26	0.662	0.305	-0.19
Consistency in Design	0.332	0.558	0.298	0.369
Search Engine Friendly	-0.55		-0.72	
Perfect Content	0.42	-0.47		0.653
Functionality	0.426	-0.54	0.312	-0.56

From the above table we can infer that the features user-friendly, perfect content, fast loading and compatibility of browsers are more significant features in classifying the websites as effective web sites.

VI. CONCLUSION

We conducted a pilot survey to understand the features which are important to the web users in usage of the web. Based on the data from the pilot survey, we conducted the final survey to collect the data using the purposive random sampling. We found the list of most significant factors which are influencing in classifying the web sites. We applied factor analysis on these factors in order to understand the most significant factors from among the list of significant factors. So we conclude that the features : user-friendly, perfect content, fast loading and compatibility of browsers are more significant features in classifying the websites as effective web sites.

REFERENCES

[1] Williams, B., Brown, T., & Onsman, A., “Exploratory factor analysis: A five-step guide for novices” , Australasian Journal of Paramedicine vol. 3 , issue 8 , 2010.

[2] Masaki Matsunaga , “How to Factor - Analyze your data right : Do 's Don'ts , and How-To's”, International Journal of psyhological research, Vol 3, issue 1, pp. 97-110 , 2010.

[3] Matt R. Raven, “The Application of Exploratory Factor Analysis in Agricultural Education Research”, Journal of Agricultural Education, Vol35, issue 4 , pp9-14, 2010.

[4] Muzammil Haque , “Sampling Methods In Social Research” ,Global Research Methodology Journal , 2010.

[5] Ji-bin Zhang, Zhi-ming Xu, Kun-liXiu, Qi-shu PanA, “Web Site Classification Approach based On Its Topological Structure” , International Journal on Asian Language Processing Vol. 20, issue 2 , pp. 75-86 , 2006.

[6] K. Selvakuberan, M. Indradevi, Dr. R. Rajaram, “Combined Feature Selection and classification – A novel approach for the categorization of web pages” , Journal of Information and Computing Science , Vol 3, Issue 2, pp. 083-089, 2008.

[7] Xiaoguang Qi , Brian D. Davison , “Web Page Classification: Features and Algorithms”, Technical Report , Lehigh University , June 2007.

[8] Charles Teddlie and Fen Yu, “Mixed Methods Sampling: A Typology With Examples”, Journal of Mixed Methods Research, vol 1, pp.77 - 99, 2007.