Assessment of Measures for Information Retrieval System Evaluation: A User-centered Approach

Bernard Ijesunor Akhigbe
Department of Computer Science & Engineering, Ile-Ife, Nigeria

Babajide Samuel Afolabi
Department of Computer Science & Engineering, Ile-Ife, Nigeria

Emmanuel Rotimi Adagunodo
Department of Computer Science & Engineering, Ile-Ife, Nigeria

ABSTRACT
The ever increasing need for Information globally is a primary reason for the scores of daily usage of IR system. Therefore there is a need to evaluate the system from a more holistic perspective – of both the system and the user. At the moment system-centered measures are not usable for the user-centered approach. Therefore, this paper attempts to determine and also suggest measures as well as methods to meet this need. The factor analytic technique was experimented for this purpose, and the structural equation modeling technique was used to estimate the resultant model. Results show that the study demonstrated high significance. Hence, the statistics presented is capable of inspiring further work in IR systems' evaluation from user's perspective.

General Terms
Evaluation, Information System, Information Retrieval, User Studies

Keywords
Structural equation modeling, Factor Analysis, IR system, Measures, System-centered paradigm and user-centered paradigm

1. INTRODUCTION

IR systems are concerned with the searching for documents; for information within documents; and for metadata about documents, as well as that of searching relational databases and the World Wide Web [39]. It is a key technology for knowledge management, which guarantees access to large corpora of both structured and unstructured data [2]. This field includes web search engines, hypertext, database and other search facilities embedded in, for example: web shops and even desktop applications [14]. Today, web search engines are even the most visible IR applications [39], hence its use in this study. It is the basic technology behind question and answering systems [15], and recently it has been introduced in the implementation of natural language processing systems [1]. It has also become an everyday technology for many web users, since it has to do with the storage and representation of knowledge and the retrieval of information relevant to a specific user problem [2] or information need. Systems with these abilities are often known as Information retrieval (IR) systems. These systems allow millions of users of the web and other applications (in which they are embedded) to express their information needs as queries, with the expectation of feedback as response to their queries. According to [2], user's queries are often compared to document representations, which are extracted during an indexing phase, with the most similar documents presented to users who are expected to evaluate their relevance with respect to their information needs and problems.

The goal of an IR system is to locate relevant documents in response to a user's query [4]. Hence, in a broad sense, it entails the process of finding information that satisfies a user's need [3], [4]. Nevertheless, IR systems and their evaluation have increased in importance. It has become a very active area of research and development with continued information explosion. This has been fueled by factors such as: The emergence of the internet; digital library initiatives; the volume of web pages in the World Wide Web and the ever increasing information need of users globally [6], [4]. Therefore the need to evaluate this system more holistically cannot be overlooked. However, IR system has been evaluated so much using the system-centered approach (SCA) [5], [6], [9], and [1], with little attention paid to the use of the user-centered approach (UCA). As a result there are standard measures [6], [11], [12], [13] only for evaluating the system from the system perspective. As buttressed by [2]; arguments are rife regarding the fact that user-oriented evaluation is extremely difficult and requires many resources. This is in view of the reason that in order to evaluate the individual aspects of searching and the subjectivity of user judgments regarding the usefulness of searches, almost an impracticable effort would be necessary. As a result IR system evaluation experiments try to evaluate only the system [2]. This is often from the system's perspective, with the user being assumed as an abstraction and not a real user. In order to achieve this, users are replaced by objective experts who judge the relevance of a document to one information need according to Mandl. This evaluation methodology is still the evaluation model for modern evaluation initiatives [2].

But these initiatives, which entails the use of the SCA to evaluating IR system has failed to address issues from user's perspective. This has brought about challenges and dearth in usable and reliable parameters for use in the UCA to evaluating IR systems. As a result there is the need to both understand and present usable measures (factors) as well as method of assessment for carrying out user-oriented studies. That there are new heterogeneous requirements and changing necessities of information needs, which must be coped with and accounted for and that there is the need for new approaches [2] to IR system evaluation, further emphasizes this need. Moreover, to achieve a system that meets the overall purpose of IR system design, which is to satisfy the need of the user [24]; its evaluation should be holistic. That is considering all the factors there is,

whether for use in the SCA or UCA, or for both. But so far this is only true of the SCA. Yet both approaches have been acknowledged and recommended in literature [3], [5], [8], [9], [11], and [12]. Thus the need to strike a balance is germane for the overall good of the system; so the system can better achieve its overall aim; and be the more accepted.

Therefore the purpose of this paper is to first propose measures that will be usable for the user-centered approach to IR system evaluation. Secondly, propose both the approach and method employed for measures assessment. Additionally, the results presented are generalizable. This is based on the fact that they are from real life IR system (search engines) and users, and were empirically tested. Section 2.0 contains the specific objectives needed to pursue the aim of the paper. Both the literature review and related works are in section 3.0. Issues that borders on methodology are discussed in section 4.0, while section 5.0 contains a lucid discourse on data analysis and results. The paper ends with a conclusion in section 6.0.

2. SPECIFIC OBJECTIVES OF STUDY

In view of the fact that the need to make available usable measures as well as method of assessment for carrying out user-oriented studies in IR system evaluation still remain an issue to be addressed; two specific objectives were pursued. This was with a view to achieving the overall aim of this paper. The objectives were to: First Identify, characterize and assess new measures from users' perspective for IR system evaluation from user's perspective; and secondly test and verify the measures identified and suggest them for use.

3. LITERATURE REVIEW AND RELATED WORKS

[7] was the first to report that the process of evaluating IR system can be classified into two: The SCA and the UCA to IR evaluation. Also, [5] reported that a total of six (6) separate levels can be gotten from the two classifications, namely: The engineering (ENG) level, the input (IPT) level, the processing (PCS) level, the output (OPT) level, the user and use (UAU) level, and the social level. While the SCA basically covers the first three levels, the UCA consists of the last three levels. Both [7], [5] argued that a paradigm shift in IR evaluation is needed from the SCA to the UCA. This has since been a challenge in the area of research on how best to evaluate IR systems. The challenge is in two fold: A shift from the system-centered paradigm orientation to the user-centered paradigm orientation of evaluating IR systems; then secondly, how to integrate results from these two paradigms, so as to achieve a better holistic approach to IR system evaluation. With this it is expected that a more comprehensive picture of IR performance would be established and as a result dangerous blind spots or gaps in research and development in the area of information retrieval (IR) would be avoided [5]. The second challenge has been well attended to in one of our papers, but in this paper the focus is primarily a contribution towards the first challenge.

The criterion of relevance as an objective, which is still been measured using precision and relevance (now known as recall) was first proposed by [17]. These metrics and objective has since become the staple of IR evaluations, even till now [5], [11], [16], [13], for use in the SCA. Other variants of precision and recall have been developed for the evaluation of IR system performance, namely: FHR (first hit success), FARR (first answer reciprocal ranking), TRR (total reciprocal ranking), Harmonic mean (a.k.a. F-measure), Novelty, the E-measure, RR (relative recall), RHL (ranked half-life), CG (cumulative gain) and the measures of (discounted) cumulated gain ((D)CG), and others [11], [12], 13]. All these measures are usable only for the SCA and have been well explored for IR system evaluation from the system perspective.

Nicholas J. Belkin [9] cited one of the proponents of IR, and said, "Already in 1988, on the occasion of receiving the ACM SIGIR Gerard Salton Award, Karen Sparck Jones argued that substantial progress in IR was likely only going to come through addressing issues associated with users (potential or actual) of IR systems, rather than continuing IR's researches almost exclusively focused on document representation and matching, and ranking techniques". This has been achieved almost a ton times and over, under the evaluation of IR systems, using recall and precision, and other related measures earlier stated in the SCA to IR system evaluation. [9] thus maintained that attention should be given to evaluating the system based on "the degree to which it benefits and support the user", as well as also examine other user related issues. As a result, this he said will bring about increase in usefulness, usability and pleasurability of IR use [9].

Despite this call, even current works such as that of: [18], [19], [20] and others, still focus on the use of the SCA for IR system evaluation. Likewise [11] also confirmed that the traditional IR system uses recall and precision to measure performance. In their work, the IR system under consideration was a web-based system. Although, there was an element of "user effort" being incorporated, it was only added as one of the evaluation criteria. This was only a tiny slice of the main criteria and as such the focus was still system-centered and not user-centered. In fact the main objective of the effort was how the system carried out its ranking processes according to a set of techniques. These techniques included the proximity of the algorithm use and the probabilistic phrase ranking technique used by the system [21].

The utility [22] of a retrieval system has to do with the difference between how much the user gained in terms of useful information, and how much the user lost in terms of time and energy. The traditional SCA to IR system evaluation lack this concept of utility. Although this was idealized in the work of [22]; first, it was not sufficiently done; and secondly the SCA was employed for the system's evaluation. With the SCA, search effectiveness [22] became the focus instead of user utility (user happiness/satisfaction). Thus the SCA status quo was still maintained.

[23] proposed a set of measures to evaluate search engine's functionality overtime. According to him, the evaluation criteria (precision and recall) and it variants used in the SCA are not sufficient. The reason being that web search engines operate in a highly dynamic, distributed environment, hence criteria and metrics so far used `are not sufficient in evaluating an IR system in a holistic manner. As a result the need for additional criteria especially to assess user utility: That is the degree of user satisfaction, which the user of a system derives as a result of use, is inevitable.

[13], in their effort investigated the quality aspects of IR system using the concept of ontology. This added complexity in terms of user interaction. However, standard IR metrics (recall and precision, and its variants) were not feasible to measure user

satisfaction (utility), which resulted from the level of interaction introduced.

[1] proposed a model of evaluation that was used to successfully assess Question and Answer System (QAS) from user's perspective. The QAS is a type of IR system that is domain specific, since it captures questions from particular domains, such as Medicine, Engineering and so on. They are meant to influence the knowledge of users. But other IR systems, like search engines do not influence the knowledge of their users, except that they assist them to locate documents and leave the user to sieve the document for relevant information, depending on their information need. Their effort yielded a 4-factor, and 18-item model for measuring user satisfaction (US) in QAS. The four primary dimensions (metric/factor) used were: ease of use, usefulness, service quality, and information quality. The limitation of the work stems from the fact that a single snapshot approach was employed; that is only one type of QAS was used, and the sample used for study was insufficient. However, the work provided some of the much needed lead in this paper.

Another related work is that of [24], where the attempt was at evaluating Information-seeking support system (ISSS). It is also a type of IR system, which focus is beyond just search and retrieval, but that found information is used to better peoples' lives and assist them toward the attainment of high-level learning objectives such as analysis, synthesis, and evaluation [41]. First [24] attended to the dilemma of how to assess Information-seeking support system (ISSS) giving the complex nature of both the system and the human environment in which they operate. This led to the second challenge, which is the limited availability of intricate metrics to assess the system. Methods such as Factor Analysis (FA) and Structural Equation Modeling (SEM) were used to assess the system using the User-centered paradigm as done in the work of [1]. This also provided another lead for this work. In the work, they used SEM to evaluate a six-factor scale of user engagement, and thus confirmed the presence of factors, such as: Aesthetics, Novelty, Involvement, Focused attention, Perceived usability and Endurability, and the predictive relationships among them. They also reported that techniques such as FA and SEM facilitate the assessment of varied, multiple, or simple measures needed for use in UCA. The work however, concluded with the assertion that if FA and SEM are used correctly, they would lead to the creation of metrics for a more holistic evaluation of ISSS. Both the efforts of [1] and [24] employed the user-centered paradigm. This paper derives from them in terms of approach in order to suggest measures for IR system evaluation using the UCA, from user's perspective.

4. METHODOLOGY

4.1 Scale items

As already highlighted the procedures employed by [1] and [24] was used in this study. These procedures are also prescribed in the work of [26] for scale development. As a result, a thorough literature search within the body of information system, which includes that of information retrieval, was conducted. This was done considering the objectives of this paper as stated in section 2. The literature search resulted in the choice of several items. This is with a view to retaining and suggesting them if certified okay using de facto standards as suggested using the work of [1], [24] and [10]. The items adopted were from previously published scales, as used in [27] and [25], but from within Information system and specifically IR system literature. These multi-items were all measured on a five point likert scale, ranging from 1 (strongly agree) to 5 (strongly disagree). In order to purify the scale a pilot study was conducted. Since, each of the items were used in previously standardized and validated scales, this was to ensure that the scale will measure exactly what it is meant for. Result of the study showed that the Cronbach alpha coefficient (CAC) range from 0.70 and above. This confirms that the scale was okay for the exercise, as affirmed by the work of [1], [27] and [10].

4.2 Data collection

The overall aim of this study was to assess measures from user's perspective, and with a view to suggest them for use in the UCA to evaluate IR systems. First, as a user-oriented study, users were identified as the main actors in the evaluation process and space. Secondly, survey method was used to collect data from those who have used one or more search engines (IR systems) according to the suggestions of [10] and [1]. The measuring instrument (MI) used was administered within and outside Nigeria. Both hardcopy of the MI and online survey method were used to elicit data from users within and outside the country respectively. A total number of 250 valid responses were used. As suggested by [28, 29], this number was sufficient to perform the necessary factor analytic (FA) process for the needed data analysis technique as required based on the objectives of the paper.

5. DATA ANALYSIS AND RESULT

5.1 Communality and eigenvalue statistics

The first statistics generated were the communality and the eigenvalue statistics. This was with a view to satisfying the first objective of this paper, which is to identify and assess new parameters in order to suggest them for use. As a result the communality value was generated to show the degree of contribution of each of the items presented based on user's assessment. By this the items selected and presented were identified in terms of their degree of contribution to assessing a component or measure. Threshold points as suggested by [30] considering their conclusion on communality values, which should be from between 0.40 to 0.70 was used. Thus the reason for the choice of > = 0.5 as the threshold point for this study. For eigenvalue result threshold of 1.0 and above was used as suggested by [1]. Below in Tables 1 and 2 are the result of communality values and eigenvalues respectively.

Table 1. Communalities Values

Communality (TP >=0.5)					
IC	**Initial**	**Extraction**	**IC**	**Initial**	**Extraction**
C1	1	.617	R1	1	.755
C2	1	.673	R2	1	.841
C3	1	.625	S1	1	.873
C4	1	.764	S2	1	.833
A1	1	.763	G1	1	.749
A2	1	.524	G2	1	.784
F1	1	.631	U1	1	.773
F2	1	.609	U2	1	.800
E1	1	.702	U3	1	.897
E2	1	.769	U4	1	.821
E3	1	.758	U5	1	.826

Communality (TP >=0.5)					
IC	Initial	Extraction	IC	Initial	Extraction
E4	1	.789			
E5	1	.654			
E6	1	.698			
T1	1	.710			
T2	1	.706			

TP (Threshold point) and IC (Item code)

Table 2. Eigenvalues before extraction

ESOSL			
C	Total	%OV	C%
1	14.926	39.278	39.278
2	3.146	8.280	47.558
3	2.226	5.858	53.416
4	1.799	4.733	58.149
5	1.522	4.005	62.154

Extraction Method: Principal Component Analysis
C (Component), %OV (% of Variance) and C% (Cumulative %)

5.2 Statistics of Factor Loadings (FLs)

The factor analytic technique was employed to extract all the items presented for this exercise, since according to the result presented in section 5.1 above, they scaled the cut off (threshold) point. Then, items that did not load strongly on any factor (values below 0.5) or had cross-loadings, were eliminated. In the work of [1], four decision rules were suggested, which are widely used and were adhered to according to [1]. The rules are: Ensure a minimum eigenvalue of 1 as a cut-off value for extraction; retain items with a factor loading (FL) of all factors greater than 0.5; assume only a simple factor structure; and for the sake of parsimony, delete factors with single-item. These rules form the bases, on which the result presented in Table 3 below was accepted. As a result a total of four (4) items were deleted. This was with a view to improve both the convergent and discriminant validity of the items presented for the study. From the Table 3 below, component 1 based on items (R1, R2, S1, S2, G1 and G2) becomes System's Dependability; component 2 usefulness base on items U1-5; component 3 ease of use base on E1-6; component 4 users' information need base on C1, T2 and component 5 system satisfaction base on C2-4 and F1.

Table 3. Summary of rotated factor loadings

C	Items	FLs (> 0.5)
1.	**R (1,2; S1,2);**	.742,.777; .803,.809;
	G (1,2)	.648, .656
2.	**U (1, 2,3,4,5)**	.752, .793, 879, .846, .840
3.	**E (1,2,3,4,5,6)**	.699, .749, .642, .706, .756, .692
4.	**C1; T2**	.502; .558
5.	**C (2,3,4); F1**	.682, .658, .773; .595

C (Components), FLs (Factor loadings)

5.3 Model's Estimation

5.3.1 Measures' and Item's Assessment

With the results of FLs presented in Table 5.2 above, a hypothesized factor structure (FS) was arrived at. The FS presents a five (5) component (measure) and 23-item model. To further ensure a non-controversial FS, CFA was used to test the model's validity and reliability. This was necessary since according to [31], EFA cannot be used to make inferences. Thus the result presented in Tables 4 and 5 below allow for proper statistical inferences. While Table 4 presents the result of the measures' (component) reliability and validity, Table 5 does for the reliability and validity of each of the individual item that form the resultant model.

Table 4. Result of model's reliability and validity 1

M	CR (> 0.6)	AVE (> 0.5)
System's Dependability	0.71	0.65
usefulness	0.55	0.71
Ease of use	0.72	0.76
Users' information need	0.80	0.63
System satisfaction	0.75	0.66

M (Measures), (CR) Composite Reliability and Average Variance Extracted (AVE)

Table 5. Result of item's reliability and validity 2

IC	FLs	IIR	IC	FLs	IIR
SD1	0.50	0.30	E2	0.75	0.56
SD2	0.68	0.46	E3	0.79	0.62
SD3	0.66	0.43	E4	0.88	0.77
SD4	0.77	0.60	E5	0.85	0.72
SD5	0.84	0.71	E6	0.84	0.71
SD6	0.59	0.35	UI1	0.71	0.50
U1	0.70	0.49	UI2	0.76	0.58
U2	0.75	0.56	SS1	0.79	0.62
U3	0.64	0.41	SS2	0.88	0.77
U4	0.71	0.50	SS3	0.66	0.43
U5	0.76	0.58	SS4	0.77	0.60
E1	0.69	0.48			

IC (Item code); FLs (factor loadings > = 0.5) and IIR (individual Item Reliability > = 0.4)

The statistics generated and used to test the reliability and validity of the model were FLs (factor loadings); Average Variance Extracted (AVE); Composite reliability (CR) and Individual item reliability (IIR). Moreover, these parameters have been used in recent studies to test the validity and reliability of models formulated using measures with their items [10], [6], [1] and [27]. Similarly, this informed their use in this study. While the statistics presented in Table 5.3.1a is the result of the assessment of the proposed measures' validity and reliability; the second result in Table 5.3.1b is the result from the

assessment of the reliability of each of the items. This was to further test the reliability of each measure in order to determine their suitability for use in IR system evaluation.

5.3.2 Model Assessment

The five (5) measures and the twenty three (23) items assessed so far forms a measurement model. This measurement model contained 23 items that describes 5 measures (components) as earlier mentioned. The model was evaluated using the structural equation modeling technique (SEM). The result presented in Table 6 below shows that that the model has a goodness of fit with the data used.

Table 6. The overall fit measurement of the model

Goodness of fit Statistics		
GOFI	**SRV**	**AV**
x^2/df	<= 3.00	2.78
GFI	>= 0.9	0.091
NFI	>= 0.9	0.097
NNFI	>= 0.9	0.093
CFI	>= 0.9	0.089
RMSR	<= 0.05	0.046
RMSEA	<= 0.08	0.066

GOFI **(Goodness of Fit Indices);** SRV **(Standard Recommended Value)** and AV **(Achieved Value);** x^2/df **(Chi square/degree of freedom),** GFI **(Goodness of fit index),** NFI **(Normed Fit Index),** NNFI **(Non-Normed Fit Index),** CFI **(Comparative Fit Index),** RMSR **(Root Mean Square Residual)** and RMSEA **(Root Mean Square Error of Approximation)**

With this result it was further confirmed that both items and measures satisfy the required standard for measures construction. Also confirmed, is the validity of the method used for both measure's and item's assessment. Since, the results arrived at is in consonance with the de facto standard under SRV column in the Table 6, then it follows that both measures and items are reliable.

6. CONCLUSION

The overall aim of this paper is to assess measures from user's perspective and with a view to proposing them for use in the user-centered paradigm for IR system evaluation. A follow up to this is to suggest both the approach and the method used for the measures assessment. This is underpinned by the fact that the measures achieved based on the result obtained and presented in this paper are the reflection of multiple items, which are dependent on user's response (assessments). Thus the only way to continuously employ these measures and ensure valid results is to use analytic methods that can handle multi-items. The factor analytic (FA) method is therefore suggested in this paper for use, in the user-centered paradigm for IR system evaluation.

Like [24] put it, the FA method; enable researchers to examine different types of user related data on user's attitudes, observed behaviors, and even system performance, and so on. [31] also reported the use of the method to evaluate a six-factor scale of user engagement; and thus confirm that measures, such as: Aesthetics, novelty, involvement, focused attention, perceived usability, and endurability, are well suited for such user related studies. Other researchers, such as [29], [32], [33], and [34] have also confirmed this suitability vis-à-vis its use in measurement model's validation. This was with a view to demonstrate that the variables (items) being measured accurately reflect the desired measures (factors), before assessing the related structural model's validity if need be. This is mostly the case when the need to assess how a measure influences others or vice versa. However, this is not one of the objectives of this paper; hence no structural model was presented.

The statistics generated as presented in section 5. above based on the methodology used, showed that the five (5) measures (System's Dependability, usefulness, Ease of use, Users' information need, system satisfaction) arrived at are valid and usable for further use in the evaluation of IR systems. In terms of validity and reliability the 5-factor and 23-item model presented in this paper hold a lot of promise in terms of significance. To ensure this, defacto standard were followed. In Tables 1 and 2 a threshold point (TP) of >= .50 and >= 1.0 were used, and the result presented showed that for all the items presented for further FA exercise were all above the standard TP, while others below TP were dropped. Also, in Table 3, a cut off point (CP) of > 0.5 was applied to arrive at the FLs presented. In the same vain, in Table 4 the TP used for CR and AVE were > 0.6 and > 0.5 respectively. Similarly in Table 5 CP and TP used were > 0.6 and > 0.5 respectively. Thus, it is clear from each of the Tables that none of the items or measures score below these cut off or threshold points. The total items presented for further FA analysis were 27. But after going through the statistical regour of the FA technique, 4 of the items were dropped. The purpose was to ensure that parsimony is avoided in keeping with the standard in literature as it is in the work of [1], [35]. Hence, the remaining 23 items were used to construct and establish the 5 new measures (System's Dependability, usefulness, Ease of use, users' information need and system satisfaction) proposed in this paper for use in the UCA. Having established the reliability and validity of the measures and individual items as shown using the result in Tables 4 and 5, it was important to estimate the model using SEM technique. The result of the resultant measurement model's goodness-of-fit presented in Figure 1 below, using SEM demonstrated a very reasonable degree of confidence that is promising and significant. While x^2/df = 2.76 and GFI = 0.931; NFI, NNFI, CFI, RMSR and RMSEA are 0.971, 0.931, 0.991, 0.041 and 0.069 respectively. All these values satisfied the benchmark recommended in literature [40], which is showed in Table 6 under SRV column. This assessment criteria has been used in some studies [25], [10], including very recent ones [36] and [37] to confirm a measurement model's validity. Thus, all constructs had strong and adequate reliability and discriminant validity.

This paper is not without some limitations. The sample size (250) employed in this paper, although satisfies the recommended size in literature [28, 29], more samples are needed. This will further strengthen the sample space, thus ensuring a wide audience of respondents. There is also the need to design evaluation exercises of the system that will include the assessment of factor (measures) that influences others based on relevant hypothesis. This is left for future work. Both items and corresponding measures suggested in this paper and the methods used for achieving them remain valid contribution to this area of research. As a result, the paper provides a possible way out of the situation described in the work of [2]. Their argument was

that users were often assumed as abstraction, and therefore excluded from the process of evaluating the IR system that was designed to satisfy (help meet) their information need, according to [38]. In conclusion, this paper is capable of inspiring as well as blaze the trail for further work in the evaluation of IR systems from user's perspective using the user-centered approach, as the result presented showed.

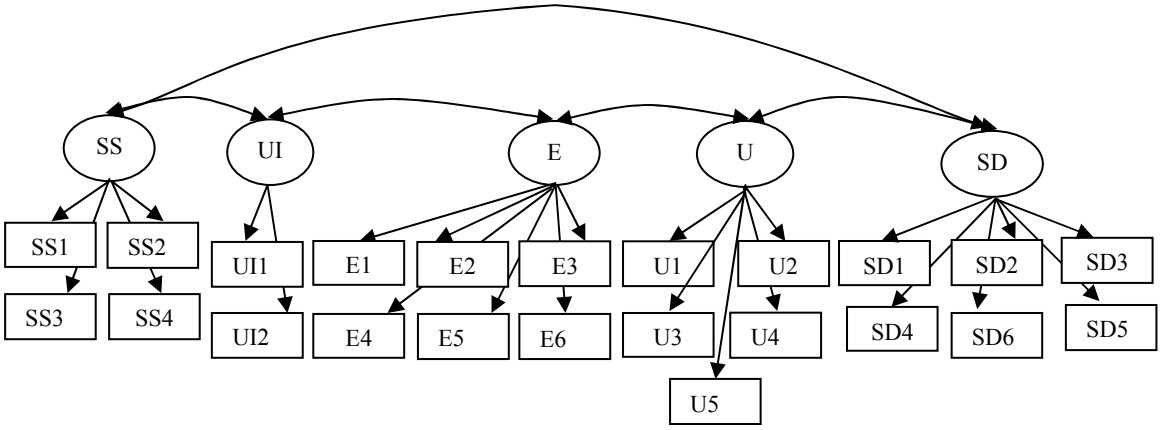

Fig 1: The measurement model and SRV: x^2/df <= 3.00; GFI >= 0.9; NFI >= 0.9; NNFI >= 0.9; CFI >= 0.9; RMSR <= 0.05 and RMSEA <= 0.08

7. ACKNOWLEDGMENTS

Several efforts resulted in this paper. To Ominiyi A. and John with his friend who handled all the technical issues that crop up regarding the online survey method employed for the study, I say a big thank you. Also to Sis Evelyn who had to part with some money to ensure that some of the hard copy questionnaires were administered, I say thank you too.

8. REFERENCES

[1] Ong, C.-S., Day, M.-Y., Hsu, W.-L. (2009). The measurement of user satisfaction with question answering systems. Elsevier, Information & Management 46 (2009), pp 397–403

[2] Mandl, T. (2008). Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance. Informatica 32 (2008) 27–38

[3] Voorhees, E., Harman, D. (2001). Overview of TREC-2001. Proceedings of TREC'2001. Available at http://trec.nist.gov

[4] Kumar, R., Suri, P.K., Chauhan, R.K. (2005). Search Engines Evaluation DESIDOC Bulletin of Information Technology , Vol. 25, No. 2, March 2005, pp. 3-10.

[5] Saracevic, T. (1995). Evaluation of evaluation in information retrieval. Proceedings of SIGIR 95, 138-46

[6] Wu M., Diane H. S. (1999). Reflections on information retrieval evaluation. In Proceedings of the 1999 EBTI, ECAI, SEER & PNC Joint Meeting. Academia Sinica accessed from http://pnclink.org/annual/annual1999/1999pdf/wu-mm on 01/03/2010 @ 12:23pm

[7] Dervin B., and Nilan M. S. (1986). Information Needs and Use. In Williams, M. E. (Ed.) Annual Review of Information Science and Technology, vol. 21, (pp.3-33). White Plains, NY: Knowledge Industry.

[8] Lewandowski, D., and Hochstotter, N. (2008). Web Searching: A Quality Measurement Perspective. A. Spink and M. Zimmer (eds.), Web Search, Springer Series in Information Science and Knowledge Management 14, pp 309- 340. Published in Springer-Verlag Berlin Heidelberg.

[9] Nicholas J. B. (2008) Some (what) Grand Challenges for Information Retrieval, European Conference on Information Retrieval (ECIR), Glasgow, Scotland, 31 March 2008.

[10] Zhilin Yang, Shaohan Cai, Zheng Zhou, Nan Zhou (2005) Development and validation of an instrument to measure user perceived service quality of information presen- ting Web portals. Elsevier, Information & Management 42 (2005) 575–589

[11] Dragomir, R. R., Hong, Q., Harris, W., and Weiguo, F. (2002). Evaluating Web-based Question Answering Systems. In Demo Section, LREC 2002, Las Palmas, Spain.

[12] Jaana K., Kalervo J. (2005) Evaluating Information Retrieval Systems under the challenges of interaction and multidimensional dynamic relevance. Proceedings of the 4th COLIS Conference. Greenwood Village, CO: Libraries Unlimited, pp. 253-270.

[13] Strasunskas, D., Tomassen, S.L. (2007) Quality Aspects in Ontology-driven Information Retrieval. In Khosrow-Pour, M. (Ed.) Managing Worldwide Operations and Communications with Information Technology (Proceedings of the 2007 IRMA International Conference), Vancouver, Canada, 2007, IDEA Group Publishing, pp. 1048-1050.

[14] Schmettow, M. (2006). User Interaction Design Patterns for Information Retrieval Systems pg (C6-1) – (C6-24) accessed @ www.hillside.net/europlop/ europlop2006 /work- shops /C6.pdf on 28/10/2008

[15] Marius Pasca and Sanda Harabagiu. High Performance question/answering. In Proceedings of the 24th International Conference on Research and Development in Information Retrieval, pages 366–374, 2001.

[16] Gao, X., Murugesan, S., and Lo, B. (2004). Multi-dimensional Evaluation of Information Retrieval Results. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04). IEEE Computer Society, pp 192- 198.

[17] Kent, A., Berry, M., Leuhrs, F. U., & Perry, J. W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation,6*(2), 93–101.

[18] Turpin A. H., and Hersh W. (2001) Why batch and user evaluations do not give the same results. In SIGIR 2001, Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 225-231). New York: ACM.

[19] Sparck, J.K. (2005). Meta-reflections on TREC. In E.M. Voorhees & D.K. Harman (Eds.) TREC: Experiment and Evaluation in Information Retrieval (pp. 421-448). Cambridge, MA: MIT Press.

[20] Turpin, A., Scholer, F. (2006). User performance versus precision measures for simple search tasks. Proc. 29th ACM SIGIR Conf., pages 11-18, Seattle, US, August 2006.

[21] Radev, D. R., Weiguo F., Hong Q., and Amardeep G., (2002). Probabilistic question answering from the web. In The Eleventh International World Wide Web Conference, Honolulu, Hawaii, May.

[22] Yiming Y., Abhimanyu L., Ni L., Abhay H., Bryan K., Monica R. (2007) Utility-based Information Distillation Over Temporally Sequenced Documents. SIGIR 2007 Proceedings, Amsterdam, The Netherlands, ACM 978-1-59593-597-7/07/0007

[23] Barllan, J. (2009). Criteria for Evaluating Information Retrieval Systems in Highly Dynamic Environments. Accessed @ http://citeseerx.ist.psu.edu/viewdoc on 04/03/09

[24] Toms, E. G., O'Brien, H. (2009). The Information-seeking Support System (ISSS) Measurement Dilemma, Published by the IEEE Computer Society, 0018-9162/09, pp. 44 -50, 2009 IEEE

[25] Wu, J.-H., Shen, W.-S., Lin, L.–M., Greenes, R., and Bates, D.W. (2008). International Journal for Quality in Health Care; Volume 20, Number 2: pp. 123–129

[26] Churchill, G. A. (1979). A Paradigm for Developing Better Measures of Marketing Constructs. Journal of Marketing Research. Vol. 16, No. 1, pp. 64-73. Published by American Marketing Association.

[27] Nauman, S., Yun, Y., Suku, S. (2009). User Acceptance of Second Life: An Extended TAM including Hedonic Consumption Behaviours. 17th European Conference on Information Systems**.** ECIS2009-0269.R1. pg 1-13.

[28] Suhr, D.D. (2005). Statistics and Data Analysis Paper 203-30 Principal Component Analysis vs. Exploratory Factor Analysis. In the Proceedings of the 30th Annual SAS Users Group International Conference. Cary, NC: SAS Institute

[29] Suhr, D.D. (2006). Statistics and Data Analysis Paper 200-31 Principal Component Analysis vs. Exploratory Factor Analysis. In the Proceedings of the 31st Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc.

[30] Costello, A.B., and Jason, W. O. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for getting the most from your Analysis. Practical Assessment Research and Evaluation, 10(7)

[31] O'Brien, H. (2008). "*Defining and Measuring Engagement in User Experiences with Technology*," doctoral dissertation, Dalhousie University, 2008.

[32] Brown, T. A. (2006). Confirmatory Factor Analysis for Applied Research. New York: Guilford.

[33] MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. Annual Review of Psychology, 51, 201–226

[34] Asparouhov, T. and Muthen, B. (2009). "Exploratory structural equation modeling". Structural Equation Modeling, 16, 397-438.

[35] Nauman, S., Yun, Y., Suku, S. (2009). User Acceptance of Second Life: An Extended TAM including Hedonic Consumption Behaviours. 17th European Conference on Information Systems**.** ECIS2009-0269.R1. pg 1-13.

[36] Byun , D. and Finnie, G. Evaluating usability, user satisfaction and intention to revisit for successful e-government websites. e-Government, an International Journal. Vol. 8, No. 1, pg 1-19. 2011

[37] Beneke, J. Towards a conceptual model: a path analysis of fundamental relationships affecting mobile advertising effectiveness. International Journal of Electronic Finance. Vol. 5, No. 1, 2011, pg 15 – 31.

[38] Al-Maskari, A., and Sanderson, M. (2010). A Review of Factors Influencing User- satisfaction in Information Retrieval. Journal of the American Society for Information Science and Technology. Published online by Wiley InterScience. Retrieved from http://disshef.ac.uk/mark /pub -licatio- ns/my_papers/2010_JASIST_Azzah. pdf on 18/03/2010 @ 1:06am.

[39] Wikipedia (2011). Information retrieval. Retrieved on 24/05/2011 @ 9:13 pm from http://en.wikipedia.org/wiki/Information_retrieval

[40] Hair, F. J., Black, W. C., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice-Hall.

[41] Ryen W. White (2009). Designing Information-Seeking Support Systems. Reports on NSF Workshop on Information Seeking Support Systems. Retrieved from http://ils.unc.edu/ISSS/ on 09/06/2011