Using semantic differentials for an evaluative view of the search engine as an interactive system

Frances Johnson
Department of Languages, Information & Communications
Manchester Metropolitan University
Geoffrey Manton
+44 161 247 6156
F.Johnson@mmu.ac.uk

ABSTRACT

In this paper, we investigate the use of semantic differentials in obtaining the evaluative view held by users of the search engine. The completed scales of bipolar adjectives were analysed to suggest the dimensions of the user judgment formed when asked to characterize a search engine. These were then used to obtain a comparative evaluation of two engines potentially offering different types of support (or assistance) during a search. We consider the value of using the semantic differential as a technique in the toolkit for assessing the user experience during information interactions in exploratory search tasks.

Categories and Subject Descriptors H3.3 [Information search and retrieval]; Search process. H.5.2 [User interfaces]: Evaluation/methodology

General Terms
Measurement, Performance, Design, Human Factors

Keywords
Semantic Differentials, User Evaluation, Exploratory Search, Information Interaction, User Interface Design,

1. INTRODUCTION

The design of interfaces to support exploratory search seeks to provide users with the tools for and the experience of an interactive and engaging search. This is a departure from the classic model of information retrieval wherein the user submits a keyword query to the system and scans the list of retrieved results for relevance, either stopping with relevant results or refining the query to get results that are closer to the information need. Exploratory search does not necessarily assume that the user has a well defined information need (at least one that can be articulated as a keyword query) or indeed that the query will be 'static' and thus satisfied by a single list of retrieved results.

Accordingly, search engine developments have focused on providing query assistance drawing on contextual aspects to the search, such as personal history and/or current context [9]. At the interface, developments focus on improving the search process via richer information representations and interactions, such as previews and facets through to tools that allow the user to view and explore connections in the results, for example 'the relation browser data analysis tool' [10]. These shifts into HCIR are intended to help in the various stages of search, from starting the task and understanding the query topic, throughout the search in deciding what to do next, and to stopping with a sense of confidence. In short, developments aim to support true exploration of the search and, whilst many efforts may fall short, they will provide some form of user support in query assistance and in improving the search process as an interactive experience.

The context for evaluation is predicated on White and Roth's [3] model of the exploratory search process. This involves the searcher in a dynamic interplay between the cognition of their 'problem space' and their exploratory activities in the iterative search process including the query formulation, results examination and information extraction. Data collected on the searcher's information interactions may confirm this model [7] as well as attempt to systematically evaluate the effectiveness of exploratory search systems. In evaluation, a framework is used to attempt to assess performance during the search stages and to relate aspects of the system to its role in supporting information exploration, including sense making or query visualisation [5]. The challenge for the evaluation of exploratory search is the assumption that the user is willing or able to make an evaluative judgment throughout the search or that valid measures can be found through their actions, for example of usage of query terms. In general, evaluation draws from established HCI measures of effectiveness (can people complete their tasks?) efficiency (how long do people take?), an assessment of the user's overall satisfaction or other affective responses. Where possible, and increasingly so, the user actions are observed and recorded as dependent on the system and/or its interface. In this study we focus on an attempt to obtain the user's evaluative view of the search engine, based on criteria which may be affected by the developments for new and richer interactive designs. It is assumed that this would be part of an assessment which when taken with others will build a picture of the 'user experience' of the system used in exploratory search.

2. USER EVALUATION

In developing an instrument to collect the user assessment effort goes into ensuring that the evaluation is made in the task context. It means little to know that the user is 'satisfied' with the interface without gaining insight into why this assessment has been formed. A variety of questionnaires have been developed for assessing usability of interactive systems, such as search engines. Two well known are the SUS (System Usability Scale) developed at the Digital Equipment Corporation [2] and the QUIS (Questionnaire for User Interaction Satisfaction) from the University of Maryland [4]. Both assess usability from the user perspective with 10 statements and rating scales in the SUS and the QUIS with 27 questions. The QUIS asks the user to respond on a rating scale to statements which address specific usability aspects of the system, such as "use of the terms were consistent throughout the website". The SUS on the other hand focuses on collecting the users' overall reaction to the site/system on statements, such as "I found the website unnecessarily complex". Arguably the QUIS focuses on the concerns that a developer might have when assessing usability whilst the SUS assumes that the user's overall assessment is a reflection on the extent to which their goal directed tasks were facilitated by the system and its design.

Questionnaires, such as SUS, are used in an experimental set up when an explanation of the user's overall assessment is sought. However, the limitations of the questionnaire to capture and provide insight into the complexity of the user's assessment has lead to alternative tools, for example Microsoft's Product Reaction Cards in the "Desirability Toolkit". This invites participants on a usability test to select as many, or as few, words from a list of 118 which best describe their reaction and/or interaction with the system they have just used. Benedek and Miner [1] includes a list of the words used and point out that the approach helps elicit negative comments as well as positive, thus overcoming a problem with questionnaires biased towards positive responses.

Given the potential scope of the users' response (represented in the reaction cards with some 100+ terms) this study sets out to investigate the value in assembling these into a framework (of sorts) for the collection of the users' evaluative judgment of an interactive system based on the technique known as 'semantic differentials'. Specifically the aim of this small preliminary investigation was to begin to determine the extent to which users hold an evaluative view of a 'search engine' and, what are the dimensions (traits or criteria) on which we form this view. If it can be found that this view is strongly held (that is, an attitude is formed which may influence how we behave and interact with the search engine) then it may be feasible to investigate the influence, if any, of a design for information interaction on the evaluative view. In this study the technique of semantic differentials is used to best describe the evaluative view held by its participants. This is then employed to assess two quite different search engines following the completion of two query based searches.

3. SEMANTIC DIFFERENTIALS

Semantic Differentials (SDs) originate from the work of Osgood [8] as a technique for attitude measurement, scaling people on their responses to adjectives in respect to a concept. Typically individuals respond to several pairs of bipolar adjectives scored on a continuum + to – and in doing so differentiate their meaning of the concept in intensity and in direction (in a 'semantic space'). The assumption made here, in the use of SDs on 'search engines' is that users hold an evaluative view which is formed when using the engine to find and/or explore information. The SD is used to investigate the adjectives that best 'conceptualise' the search engine, from the user perspective. Factorial analysis is also used to identify the dimensions of the judgment, in a sense the packaging of the components of the judgment into smaller units of meaning reflecting what is important when responding to the concept 'search engine'.

The design of the SD aims to allow a degree of abstraction in the evaluation so that participants can reflect the complexity of their response. In this study, the adjectives to include on the SD scale were chosen from Microsoft's Product Reaction Cards, these having been collected in previous research, usability studies and in the marketing of web sites and systems. The majority of the terms formed pairs on some continuum and 40 terms (20 pairs) were selected to present in the SD. The selection was subject to the judgment of the researcher. This is a limitation of this exploratory study, however some steps were taken to formalise the selection. A loose grouping of the adjective pairs was made as relating to appearance (such as 'attractive'), judgment ('relevant'), emotive ('boring') and use ('fast'). Five pairs from each of these groupings were made. The pairs were mixed on the SD to avoid having all the positive terms on one side of the scale and only intervals were shown on the scales with the numerical values used only for data entry. This allowed participants to focus on how an adjective pair related to the engine and its characteristics, rather than on 'scoring' it in some way.

3.1 Implementation

The study was conducted on our undergraduates studying BSc Web Development and on a postgraduate cohort studying on MA Library and Information Management or the MSc Information Management. A total of 89 students participated in the study. At the start of the class each participant was asked to think about a search engine, and adjectives they would use to describe the engine, (in other words, "what it means to them"). Each participant was then given the SD to complete. This is referred to as the 'baseline' and the data were analysed to gauge user perceptions of search engines.

In the following lab sessions (about one hour later) each participant was required to perform two search tasks on each of the two search engines - Google, an engine we can assume some familiarity and, a second clustering engine (Yippy, formerly Clusty). The two tasks were as follows

1. Find information on the symptoms for diabetes type II
2. Find information to help write an assignment on the debate 'nurture vs nature'

These were selected to give the participants experience of using the engines for a closed question (find symptoms) and on a more open 'informational' type of query (on the 'nature nurture' debate). A measure of search success was not taken as the aim was simply to get the participants using the engines. The order of use of the two sites was randomized so that approximately half of the participants worked on Google first and half on the clustering engine. All were told to spend no longer than 10 minutes searching on each engine and to complete the SD for each engine immediately after each use.

4. FINDINGS

4.1 Evaluative views

The responses to the baseline (*think of an engine*) were entered into SPSS with the scales coded (7-1) so that the positive adjectives corresponded to the higher numbers. Descriptive statistics of mean, mode and standard deviation were calculated for each of the adjectives. Those with a mean greater than 4 or less than 3 were taken to suggest the adjective pairs that best characterise the participants' view, as follows

attractive	unattractive
powerful	simplistic
valuable	not valuable
relevant	irrelevant
satisfying	frustrating
fast	slow
predictable	unpredictable
intuitive	rigid
easy	difficult

Factor analysis investigates the correlations among subsets of the responses to the bipolar pairs and groups the correlated variables such that each group is largely independent of the others. Exploratory factor analysis was employed to identify the groups which might explain most of the variance in the data. With 20 pairs of adjectives to perform Principal Components Analysis (PCA) in SPSS it is recommended that a minimum of 100 responses are obtained, whilst others recommend that a sample requires approx 5-10 times the number of people as scale pairs [6]. With 89 responses we should use a reduced number of pairs, however the Kaiser-Meyer-Olkin measure of sampling adequacy (.616) is greater than the 0.6 needed to indicate that the correlations matrix may be able to factorise. So with this, PCA was run (with varimax rotation to force items to 'load' with only one factor group), to identify the possible 'factors' or subsets derived from patterns of correlation of the adjective pairs. The following five subsets were obtained (the adjectives from the list above having a low or high mean are shown in bold). The labels were assigned to suggest the evaluative dimension.

Factor 1 *label* USE – Utility

effective, **valuable, satisfy**, **relevant**, **predictable**, intimidating, inspiring, stimulating

Factor 2 *label* QUALITY – Affective

engaging, fun, connected

Factor *3 label* QUALITY - Appearance

high quality, personal, meaningful, **rigid, attractive**

Factor 4 *label* USE – Efficient

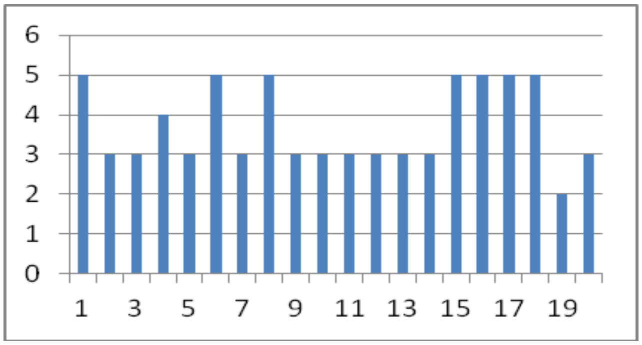
easy, intuitive, fast, powerful

Factor 5 *label* USE - Control

controllable

4.2 Comparative evaluations

Using the same SDs, participants scaled their responses post search using Google and the clustering search engine. These were entered into a worksheet to obtain basic statistics. The mode for each adjective is shown Figure 1 with a note of those with mode >4 and < 3 suggesting a positive or negative response.



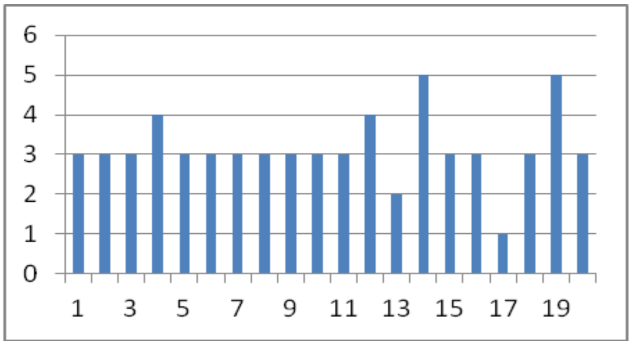
Google (mode > 4 or < 3)

& in bold where mean is also > 4

1**attractive** - , 6valuable - , 8relevant - ,

15**satisfying** - , 16**fast** - , 17predictable - ,18**controllable** -,

and (where mode < 3) 19rigid -



Clustering search engine (mode > 4 or < 3)

& in bold where mean > 4 or < 3

14engaging - , 19intuitive –

and (where mode < 3)

13intimidating - , 17 – **unpredictable**

Figure 1. Responses to the adjectives for both engines

Using the suggested dimensions or aspects of the user evaluation from the factor analysis of the 'baseline' data we can compare the participants' responses on the high or low scoring adjectives across the engines. On QUALITY – Appearance Google was rated rigid and attractive and whereas Google was neutral on the factor QUALITY- Affective, the clustering search engine obtained a positive score towards the adjective engaging. On the factor labeled USE- Utility Google was scored as predictable, valuable, relevant and satisfying, whereas the clustering engine as unpredictable and towards intimidating. On USE-Efficient Google was rated as fast and the clustering engine appears more intuitive. Google was also rated as controllable.

5. DISCUSSION

This is an exploratory study and it has its limitations. It is questionable whether the selection of the adjectives to use in the SD influenced the results. In particular there is uncertainty in the results that *intuitive to rigid* is on some continuum. Also there is some unease at accepting a factor with 8 out of 20 pairs and one with only one. Perhaps the sample size was too small to attempt factoring. The results also raise questions on how some of the adjectives were interpreted by the participants. These withstanding, the participants in this study did appear to hold an evaluative judgment of the concept 'search engine' and the traits represented in the scale were grouped to suggest the aspects on which an assessment may be formed. It is of particular interest that upon using the search engine Google to conduct a search task the ratings on the SD, on the whole, altered only in the factors of 'controllable' and USE -efficient (easy, intuitive and powerful). Perhaps we can assume that Google was the typical engine when asked to think of an engine in the baseline and, when it came to *use* Google, users shifted their perception with regards to some of the adjectives. Perhaps this is not surprising but it may suggest that we hold an implicit view of search engines, and that this view will be influenced by actual use (and the experience). Our participants may have had less familiarity with the clustering engine, and in the evaluation this appears to have prompted an 'affective' response in finding the engine to be 'engaging' whilst also indicating shifts in the 'use' factors (towards an assessment of the engine as 'unpredictable'). Again the infallibility of some of the terms is highlighted where an 'unpredictable' system may be regarded to be a negative judgment, but if the system is also considered to be engaging the assessment could be highly desirable depending on the user's goals. This study of the use of semantic differentials indicates that it is worth running the test with a new cohort of students to determine the extent to which a consistent view is obtained. As an exploratory study it also suggests that further research on user's perceptions and mental models of search engines is worthwhile. With regards to the challenge of providing an evaluation of the exploratory search, this study falls short as no behavioural data was obtained. However, perhaps, with further design of the SD and use in an experimental set up with honed tasks, a user assessment of the interface may be obtained as dependent on the search interface development and design.

6. REFERENCES

[1] Benedek, J. and Miner, T. "*Measuring Desirability: New Methods for Evaluating Desirability in a Usability Lab Setting*." Redmond, WA: Microsoft Corporation, 2002. http://www.microsoft.com/usability/UEPostings/Desirability Toolkit.doc

[2] Brooke, J. SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), *Usability Evaluation in Industry*. London: Taylor & Francis, 1996 [*www.itu.dk/courses/U/E2005/litteratur/***sus***.pdf#*

[3] Capra, R., and Marchionini, G. The Relation Browser tool for faceted exloratory search. Proceedings of the 2008 Conference on Digital Libraries, Pittsburg, Pennsylvania, June, 2008

[4] Chin, J. P., Diehl, V. A, & Norman, K. Development of an instrument measuring user satisfaction of the human-computer interface, Proceedings of ACM SIGCHI ,1988, pp. 213-218. http://www.cs.umd.edu/hcil/quis/

[5] Daqing He, et al An evaluation of adaptive filtering in the context of realistic task-based information exploration. . *Information Processing Management*, 44(2), 2008 pp. 511-533

[6] Gable, R. K., & Wolf, M. E.. *Instrument development in the affective domain* (2nd ed.). Boston: Kluwer Academic, 1993

[7] Kules, B and Capra, R. Visualizing stages during an exploratory search. Proceedings HCIR October 20th, 2011.

[8] Osgood, C.E, Suci, G., & Tannenbaum, *P The Measurement of Meaning*. University of Illinois Press, 1957

[9] Teevan, J., Dumais,S.T and E. Horvitz. *Potential for Personalization*. ACM Transactions on Computer-Human Interaction special issue on Data Mining for Understanding User Needs, 17(1), 2010 http://people.csail.mit.edu/teevan/work/ publications/ papers/tochi10.pdf

[10] White, Ryen W. & Roth., R. A. *Exploratory Search: Beyond the Query-Response Paradigm*, CA: Morgan and Claypool, 2009

Appendix: The Semantic Differential scale

attractive	_ _ _ _ _ _ _	unattractive
impersonal	_ _ _ _ _ _ _	personal
dull	_ _ _ _ _ _ _	fun
powerful	_ _ _ _ _ _ _	simplistic
disconnected	_ _ _ _ _ _ _	connected
valuable	_ _ _ _ _ _ _	not valuable
high quality	_ _ _ _ _ _ _	low quality
irrelevant	_ _ _ _ _ _ _	relevant
effective	_ _ _ _ _ _ _	ineffective
incomprehensible	_ _ _ _ _ _ _	meaningful
stimulating	_ _ _ _ _ _ _	confusing
boring	_ _ _ _ _ _ _	inspiring
intimidating	_ _ _ _ _ _ _	empowering
stressful	_ _ _ _ _ _ _	engaging
satisfying	_ _ _ _ _ _ _	frustrating
fast	_ _ _ _ _ _ _	slow
predictable	_ _ _ _ _ _ _	unpredictable
controllable	_ _ _ _ _ _ _	uncontrollable
intuitive	_ _ _ _ _ _ _	rigid
difficult	_ _ _ _ _ _ _	eas