A Meta-Analysis of the Effects of Search Experience on Search Performance in Terms of the Recall Measure in Controlled IR User Experiments

Ying-Hsang Liu
School of Information Studies
Charles Sturt University
Wagga Wagga NSW 2678 Australia
yingliu@csu.edu.au

Abstract

This paper reports a meta-analysis of the effects of search experience on search performance in terms of the recall measure in controlled IR user experiments. More specifically, this study was designed to answer the research question: how large is the average effect size in the set of studies included in the meta-analysis? Search experience, a manifestation of users' search skills accumulated through their interactions with IR systems over time, has been identified as an important research variable in user search behaviours. The participants included in primary studies were end-users or intermediaries recruited for IR user experiments. The results of the meta-analysis (N = 8) using a fixed-effects model showed that search experience has an overall positive effect on the recall measure (weighted mean correlation coefficient r = 0.04, 95% confidence interval was -0.01 to 0.09). Our findings may provide implications for designing adaptive or personalized IR systems that take into account the contextual information at the user and interactional levels.

Keywords

Information Retrieval, User Studies Involving Documents

1 Introduction

Search experience has been identified as one of the key user characteristics that affect search performance in information retrieval (IR) user experiments (see e.g., [13, 14]). While the search experience as an important research variable has been operationalized in various ways for research purposes, search experience in general is a manifestation of users' search skills accumulated through their interactions with IR systems over time.

Proceedings of the 15th Australasian Document Computing Symposium, Melbourne, Australia, 10 December 2010.

Previous studies that were conducted in the 1980s and early 1990s revealed that end-users usually had limited experiences searching online bibliographic databases, because online searching was very expensive and professional librarians usually conducted the search on behalf of users. Here search experience usually referred to whether searchers have had extensive use of online databases and whether they were proficient in the system features, such as search commands or indexing thesauri.

For example, the search experience was measured by the total number of searching sessions in a longitudinal study of medical students' use of MEDLINE [16]. Several studies that examine the effect of search experience on searching behaviour have used the total time spent using a particular online database or DIALOG system as a measure of different levels of search experience [6, 11]. In other studies that investigated whether search success depends on searchers' individual characteristics, the search experience was determined by formal training in online database searching [1, 20].

More recent studies tend to assess whether the search experience in a specific type of information retrieval system can be transferred to another. For instance, since one of the primary objectives was to investigate the effect of online database search experience on Web search performance, researchers used the duration and frequency of using online databases to measure undergraduate students' search experience [15]. Because of the similar system features in Boolean logic, researchers used the frequency of online public access catalogue as a measure of undergraduate students' search experience in a Boolean-based online database [23].

Overall, the participants included in these studies were end-users or intermediaries who conducted searches on behalf of users, and their different levels of search experiences were measured by formal training in online database searching or various kinds of indicators of their exposure to IR systems.

The choice of performance measures of precision and recall has been widely used in evaluating the

effectiveness of automatic indexing techniques, in part because researchers can test the performance of different retrieval techniques in a laboratory environment. While user-oriented measures, such as user satisfaction and utility, have been proposed as measures of user search performance, the precision and recall measures still dominate IR experimentation research. We particularly considered the recall measure as dependent variable since it has also been extensively used in previous IR user experiments, and several researchers hypothesized that search experience is correlated with the search outcome in terms of the recall measure [6, 11, 21].

Our review of related studies have focused on controlled IR user experiments because they have high levels of internal validity and allow us to examine the subtle effects of individual differences on search performance in laboratory settings.

Despite different measurement in these user studies, the study of the impact of search experience on search performance has had a growing body of research (See [14] for a recent review). One of the outstanding questions is whether searchers' individual characteristics, such as search experience, are correlated with the measures of search performance? If the answer is yes, how can we estimate the effect of search experience on search performance?

To advance our understanding of the impact of search experience on search performance, this study was designed to collect, analyse and synthesize the empirical findings from controlled IR user experiments. The results will not only help us better understand the impact of individual differences on search performance, but also provide implications for designing adaptive or personalized IR systems that take into account the contextual information at the user and interactional levels.

We conducted a quantitative review of empirical studies by comparing and synthesizing separate results from the research literature. The technique of meta-analysis allowed us to synthesize the research results and determine the relationships between variables. More specifically, our research question is: *how large is the average effect size in the set of studies included in the meta-analysis?* In view of previous research, we formulate the following research hypothesis: *Experienced searchers will perform better than novice searchers in terms of the recall measure*.

2 Method

To collect the empirical controlled IR user experimental studies, we conducted a comprehensive search of Web of Science databases, specifically Social Science Citation Index SSCI) and Science Citation Index (SCI) in August 2008. By using the citation pearl growing search strategy [8], which was designed to use citation relationships to find relevant articles, we were able to systematically collect eligible studies for inclusion in the review.

Originally we had four pearl (or seed) articles drawing from the researcher's knowledge: Pao and her colleagues [16], Howard [11], Fenichel [6] and Sutcliffe, Ennis and Watkinson [21]. The reviewed articles in the dataset of [14] were also included as seed articles because they contain some potentially relevant studies. Using the cited reference function, with particular attention to name variants and inconsistencies of citations, our searches yielded a total of 537 unique references. The study eligibility criteria were controlled IR user experiments that involved the variables of search experience and search performance in terms of the recall measure. The researcher examined the title and abstract of each bibliographic record. Full-text of the articles were consulted if the study has a good chance of fulfilling the above mentioned eligibility criteria. A total of 104 full-text articles were examined. Our selection only resulted in two definitely relevant articles; another three was collected from an examination of cited references in the articles.

For descriptive purposes, each study was coded by searcher characteristics, sample size, IR system used, test collection, search task and outcome measure (See Appendix 1). Note that most studies used Boolean-based IR systems for experimental purposes, and the experimentation of retrieval techniques was not the primary objectives.

To measure the strength of the linear relationship between two variables, i.e., search experience and search performance, we selected correlation coefficient r. The effect size of these studies was transformed into raw correlation coefficients because the search experience variable was measured in a wide variety of ways and the outcome variable of recall was applied in different ways (See Appendiex 1). In these situations regression coefficents are not directly comparable across all the studies, while correlation coefficients can be compared [19].

Correlation coefficients of included studies were calculated based on the experimental design, sample size and details of reported statistics, using the formula in Borenstein [2] and the functions in R statistical software [5, 18]. In general, correlation coefficients can be easily computed if the report provides F value for one-way ANOVA in comparing two groups. When the F value was not available and the raw data was presented in the report, a one-way ANOVA was conducted (See [21]). For repeated measure study, such as [9], we followed the procedure in [19]. In other cases where the F value or p-value of insignificant results was not reported, the effect size was replaced with a value of zero [17], including studies of Fenichel [6], Howard [11] and Pao et al. [16].

To estimate the magnitude of search experience on search performance in terms of the recall measure, we fit the data into a fixed-effects model [12]. We assumed that there is true effect of search experience across all the studies. After deriving the raw correlation coefficient, we performed the Fisher's *r*-to-*z* transformation for normalization. The meta-analysis with a fixed-effects model was conducted using metafor package [18, 24].

3 Results

This study was designed to integrate studies that investigated the impact of search experience on search performance in terms of the recall measure in controlled IR user experiments. After the systematic collection and examination of potentially relevant articles, our corpus consists of 9 studies.

To test whether the true effect is homogeneous, a test for homogeneity revealed that homogeneity of correlations is rejected (Q = 68.09, $df = 8$, $p < .0001$). We then calculated leave-one-out diagnostics that indicates the effect of deleting one case on the fitted model [7, 24]. The results indicated that the amount of heterogeneity is significantly reduced by removing Hersh and Hickam's [9] study (Q= 7.52, $df = 7$, $p = 0.38$). Further examination of this study showed that methodologically it is different from other included studies because of the use of replicated searches for comparing search performance between librarians and physicians. Therefore, our final results were based on a corpus of 8 studies, excluding the out-lying case.

To gauge the size of homogeneity, the I^2 statistic was calculated [10]. The $I^2 = 6.9\%$ was considered small heterogeneity, suggesting that only about 7% of variation in effect sizes is due to heterogeneity

Results of the meta-analysis ($N = 8$) showed that search experience has an overall positive effect on the recall measure (weighted mean correlation coefficient $r = 0.04$, 95% confidence interval was -0.01 to 0.09), as shown in Figure 1. The figure indicated that only Chen's [4] study has demonstrated significantly positive effect of search experience, as the lower bound of confidence interval (CI) does not cross the vertical line, with zero Fisher's *z* transformed correlation coefficient. The sizes of rectangular represent sample size for each study, whereas the diamond summarizes the averaged effect size. Because the average effect size $r = 0.04$ and 95% CI was between -0.01 and 0.09, our hypothesis that experienced searchers will perform better than novice searchers in terms of the recall measure was not supported.

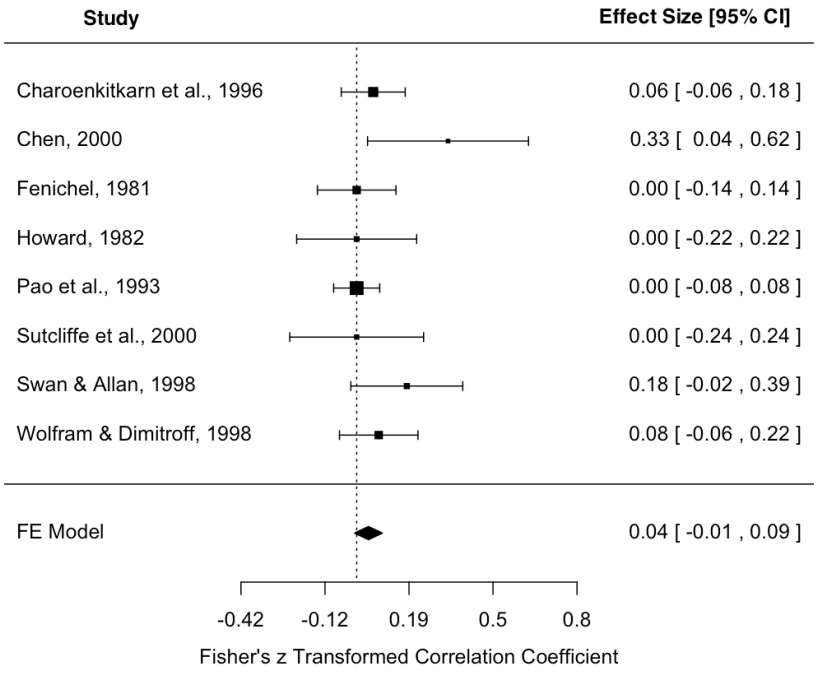

Figure 1. Forest plot of the effect of search experience on search performance in terms of the recall measure in controlled IR user experiments.

4 Conclusion

This meta-analytic study was designed to estimate the effect of search experience on search performance in terms of the recall measure in controlled IR user experiments. Our results (N = 8) indicated that search experience overall has an overall positive effect on the recall measure (weighted mean correlation coefficient r = 0.04, 95% confidence interval was -0.01 to 0.09). However, the hypothesis that experienced searchers will perform better than novice searchers in terms of the recall measure was not supported.

5 References

[1] Bellardo, T. An investigation of online searcher traits and their relationship to search outcome. *Journal of the American Society for Information Science*, 36, 4 1985), 241-250.

[2] Borenstein, M. *Effect sizes for continuous data*. Russell Sage Foundation, New York, 2009.

[3] Charoenkitkarn, N., Chignell, M. and Golovchinsky, G. Is recall relevant? An analysis of how user interface conditions affect strategies and performance in large scale text retrieval. *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*1997), 211-232.

[4] Chen, C. Individual differences in a spatial-semantic virtual environment. *Journal of the American Society for Information Science*, 51, 6 2000), 529-542.

[5] Del Re, A. *compute.es: Compute effect sizes*. R package, 2010.

[6] Fenichel, C. H. Online searching: Measures that discriminate among users with different types of experiences. *Journal of the American Society for Information Science*, 32, 1 (Jan 1981), 23-32.

[7] Greenhouse, J. B. *Sensitivity analysis and diagnostics*. Russell Ssage Foundation, New York, 2009.

[8] Harter, S. P. *Online information retrieval: Concepts, principles, and techniques*. Academic Press, Orlando, FL, 1986.

[9] Hersh, W. and Hickam, D. Use of a multi-application computer workstation in a clinical setting. *Bulletin of the Medical Library Association*, 82, 4 1994), 382-389.

[10] Higgins, J. P. T., Thompson, S. G., Deeks, J. J. and Altman, D. G. Measuring inconsistency in meta-analyses. *Br. Med. J.*, 327, 7414 (Sep 2003), 557-560.

[11] Howard, H. Measures that discriminate among online searchers with different training and experience. *Online Review*, 6, 4 (Aug 1982), 315-327.

[12] Konstantopoulos, S. and Hedges, L. V. *Analyzing effect sizes: Fixed effects models*. Russell Sage Foundation, New York, 2009.

[13] Meadow, C. T., Marchionini, G. and Cherry, J. M. Speculations on the measurement and use of user characteristics in information retrieval experimentation. *Canadian Journal of Information and Library Science*, 19, 4 1994), 1-22.

[14] Moore, J. L., Erdelez, S. and Wu, H. The search experience variable in information behavior research. *Journal of the American Society for Information Science and Technology*, 58, 10 2007), 1529-1546.

[15] Palmquist, R. A. and Kim, K. S. Cognitive style and on-line database search experience as predictors of Web search performance. *Journal of the American Society for Information Science*, 51, 6 (Apr 2000), 558-566.

[16] Pao, M. L., Grefsheim, S. F., Barclay, M. L., Woolliscroft, J. O., McQuillan, M. and Shipman, B. L. Factors affecting students use of MEDLINE. *Computers and Biomedical Research*, 26, 6 (Dec 1993), 541-555.

[17] Pigott, T. D. *Hadling missing data*. Russell Sage Foundation, New York, 2009.

[18] R Development Core Team *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, 2010.

[19] Rosenthal, R., Rosnow, R. L. and Rubin, D. B. *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press, New York, 2000.

[20] Saracevic, T., Kantor, P., Chamis, A. Y. and Trivison, D. A study of information seeking and retrieving: I. Background and methodology. *Journal of the American Society for Information Science*, 39, 3 1988), 161-176.

[21] Sutcliffe, A. G., Ennis, M. and Watkinson, S. J. Empirical studies of end-user information searching. *Journal of the American Society for Information Science*, 51, 13 2000), 1211-1231.

[22] Swan, R. C. and Allan, J. Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21st ACM SIGIR Conference* (Melbourne, Australia, 1998). ACM, New York.

[23] Vakkari, P., Pennanen, M. and Serola, S. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39, 3 2003), 445-465.

[24] Viechtbauer, W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36, 3 2010), 1-48.

[25] Wolfram, D., Volz, A. and Dimitroff, A. The effect of linkage structure on retrieval performance in a hypertext-based bibliographic retrieval system. *Information Processing & Management*, 32, 5 1996), 529-541.

Appendix 1. Descriptive analysis of the effect of search experience on search performance

Study	User	Sample Size *N*	IR System	Collection	Search Task	Outcome Measure
1. Charoenkitkarn et al., 1996 [3]	Most experienced searchers had extensive online searching experiences, and performed searches on a daily basis.	36 searchers × 8 topics = 288 searches	Information exploration system, with different search interface conditions	TREC-3 test documents	Find answers to search topics; 8 search topics from TREC-3	Standard recall
2. Chen, 2000 [4]	The average online search experience was 5 years.	12 searchers × 4 topics = 48 searches	Information visualization system, with textual and spatial search interfaces	169 articles from ACM CHI conference proceedings	4 search topics; save relevant articles for each topic	LSI (Latent Semantic Indexing)-based recall scores
3. Fenichel, 1981 [6]	Experienced searchers were regular users of DIALOG and novice searchers were beginning MLIS students.	48 searchers × 4 topics = 192 searches	ERIC ONTAP on the DIALOG system, command line interface	35,000 bibliographic references, about 12% of the ERIC database	4 search topics	Standard recall
4. Hersh & Hickam, 1994 [9]	Experienced searchers in comparison were medical reference librarians and physicians	4 times searched × 106 topics = 424 searches	GRATEFUL MED and ELHILL search interfaces	A subset of MEDLINE covering 270 journals over five years	106 search topics	Standard recall
5. Howard, 1982 [11]	Search experience was distinguished by the length, number of frequency of searches, and ERIC use experience	42 searchers × 2 topics = 84 searches	DIALOG system, command line interface	ERIC database	2 search topics	Standard recall
6. Pao et al., 1993 [16]	Medical students' search experience was based on the total number of online sessions	184 searchers × 3 topics = 552 searches	PaperChase search interface	MEDLINE database	3 search topics	Standard recall
7. Sutcliffe et al., 2000 [21]	Medical students' search	17 searchers	WinSPIRS search	MEDLINE database	4 search topics	Standard recall

	experience was based on whether they had some experience using MEDLINE	× 4 topics = 68 searches	interface			
8. Swan & Allan, 1998 [22]	Experienced searchers were librarians who had MLS degrees; Novice searchers were primarily students	16 searchers × 6 topics = 96 searches	Inquery search engine with three different search interfaces	A subset of the TREC collection; articles from the Final Times, approximately 200,000 articles	6 search topics; Identify as many aspects of relevance to a query as one can	Aspectual recall
9. Wolfram & Dimitroff, 1998 [25]	Search experience was based on searchers' self rating	48 searchers × 4 topics = 192 searches	A prototype hypertext system and a Boolean-based system	Approximately 3,000 records from the NTIS database	4 search topics	Standard recall