Rethinking the Cost of Information Search Behavior

Yinglong Zhang
School of Information and Library Science
University of North Carolina at Chapel Hill
216 Lenoir Dr, Chapel Hill, NC 27599
yinglongz@unc.edu

Jacek Gwizdka
School of Information
University of Texas, Austin
1616 Guadalupe St. Austin, TX 78703 USA
sigir2016@gwizdka.com

ABSTRACT

In this paper, we present a cognitive-economic approach to examining the cost in information search. Unlike previous studies on economic models, we calculated the cost in information search based on participants' eye-tracking data as well as their behavioral data, such as query formulation, search task duration, SERP and web page visits. Using Principal Component Analysis (PCA), we explored a possible latent factor structure of variables representing the cost in information search. Our results indicated that the cost of information seeking could be associated with two distinct aspects of search, exploratory and validation processes.

Keywords

Information seeking stopping behavior, eye tracking.

1. INTRODUCTION

An extensive review of prior research on information search has been made into how people define information needs, generate queries, examine documents, refine queries, and make sense of the information they find (for a review see [1]). It has been found that terminating search too early or too late can bring harmful effects to its outcomes [2, 3]. In previous studies, information search stopping behavior has been investigated from different perspectives. Based on the economic principle as well as the findings reported by Azzopardi and his collegues [4-6], the cost of information search is one of the most important factors contributing to stopping behavior. Inspired by theories and methods from cognitive science, we report on a cognitive-economic approach to calculating the cost in information searching behavior.

2. RELATED WORK

Economic principle is one of the well-developed theories for understanding stopping behavior in information search. This principle was first proposed by Varian [7]. In his SIGIR 1999 keynote, he addressed the relation between economics and search, and pointed out the feasibility of applying economics to solve several questions in Information Retrieval (IR), such as how to examine the economic value of information based on consumer theory, how to better estimate the probability of relevance, and how to apply Stigler's theory [8] on Optimal Search Behavior to IR.

An integral part of the economic perspective is to estimate the gain and the cost in information search. In previous studies, the cost has been mainly considered to result from a sequence of interactions. For instance, Azzopardi and his colleagues [4-6] simulated the cost of search in terms of a number of queries posted (Q), a number of search result pages examined per query (V), a number of snippets inspected per query (S), and a number of documents assessed per query (A) with some probability $\boldsymbol{P_a}$. The total of interaction was modeled as shown in **Eq 1**:

$$c(Q,V,S,A) = c_q.Q + c_v.V.Q + c_s.S.Q + c_a.A.Q \quad (1)$$

Where c_q is the cost of a query, c_v is the cost of viewing a page, c_s is the cost of inspecting a snippet, and c_a is the cost of assessing a document.

In contrast to the models proposed in [4-6], Smucker proposed that the gain and the cost can be expressed in terms of time [1]. Specifically, if users invest time ***t*** in assessing a result list, they will have a total benefit of ***G(t)*** - the cumulative gain experienced by the users at time ***t***. Assuming a decay function ***D(t)***, in this time-based model, users will stop searching information when ***D(t)*** decreases to 0. The time-based model is shown in **Eq 2**:

$$\frac{1}{N}\sum_{k=1}^{\infty} g_k D(T(k)) \quad (2)$$

Where *T(k)* refers to the expected time it would take for a user to reach result rank *k* and begin to judge the document.

Although the aforementioned studies investigated the cost of search based on varying perspectives, all of them essentially considered information search as one-stage process. However, a number of well established information search models include multiple stages [9, 10]. Thus, it is possible that some costs are only associated with a specific stage of information search and that the cost associated with different stages can have varying effects on information search behavior. In addition, the impact of individual differences on information search was neglected by most of these studies. For instance, when reading the same irrelevant document, participant A may invest more cognitive effort than participant B due to the differences in their working memory capacity. Lastly, in Azzopardi's models, all of the queries were regarded as the cost. However, it has been found that generating queries of high quality also brings benefit to information search process [11-13], though all queries carry a cost associated with them as it takes effort to generate and type them. In this regard, it could be more proper to only include queries with low quality in the estimation of the cost in information search.

3. PROPOSED MEASURE OF THE COST OF INFORMATION SEARCH

In our research, we have developed a method that includes both economic and cognitive aspects of information searching behavior. Specifically, we add eye fixations to the assessment of the cost in information search. Previous research has demonstrated that the internal cognitive activity is reflected, to some extent, in the change of the number of fixations [14, 15]. In our study, the number of fixations was used to reflect the

SIGIR'16, July 17–21, 2016, Pisa, Italy.
 ISBN 978-1-4503-4069-4/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914742

cognitive effort that each participant invested when they examine documents and search engine result pages (SERPs). We consider the cost of information search results from query generation, document examination, search engine result pages (SERPs) and task description examinations.

3.1 Cost in Query Generation

Recall that generating queries of high quality can bring benefit to information search process [11-13], though all queries carry a cost with them as it takes effort to generate and type them. In our work, only *not useful* queries were considered in assessing the cost associated with query generation; their cost was reflected in two variables:

- The number of words used in all *not useful* queries $W_{notUseful}$
- Total time taken to enter *not useful* queries $T_{notUseful}$

***Not useful* queries** were defined here as the queries that were used only once and did not return relevant documents (as judged by participants). For the queries that returned no relevant documents, but were used more once, we made an assumption that participants re-typed them on purpose. That is, these queries are not treated here as cost because they likely provided information that aided the participants in information search, for example, in reformulating existing or creating new queries. Queries that returned relevant documents are not treated as cost, because they directly facilitated discovery of the relevant documents.

3.2 Cost in Document Examination

In information search, users might invest time and effort in reading some documents that are not useful for meeting their search goals. Examining such *not useful* documents could be regarded as a kind of a cost. This cost was reflected in the variable:

- The number of fixations on visited *not useful* documents $D_{notUseful}$

***Not useful* documents** were defined as the examined documents that were located through *not useful* queries. This definition was based on an assumption that the documents returned by a query were less likely to be located by another completely different query (word stems of terms in both queries were different) in our experiment. That is to say, all of the assessed documents, located by *not useful* queries, were considered not useful, since useful documents could either be saved by participants as a relevant document (the interfaces in our experiment supported this functionality) or be found again by re-typing certain queries and then be re-examined. In contrast to *not useful* documents, all of the other documents that participants examined in the tasks were considered *useful* documents. The documents that were not examined were not considered as a cost since participants did not invest any effort reading them.

3.3 Cost in SERP and Task Description Examinations

The costs of examining SERPs were estimated based on the variable:

- The number of fixations on all SERPs F_{SERP}.

Task descriptions were available to participants during their searches in our study. We assume that participants opened task description, when they forgot or were not confident about information they were asked to find and needed to confirm it. In our study, opening tasks descriptions was regarded as a cost in the information search process. The cost of description examination was modeled, as follows:

- Total time of visiting task descriptions $T_{taskDescription}$
- The number of fixations on visited task descriptions $F_{taskDescription}$

4. METHOD

We conducted a controlled, lab-based experiment to investigate Web searches on Wikipedia. Each experiment session was held in Information eXperience lab at the School of Information, University of Texas at Austin .

4.1 Participants

32 university students (15 females), aged 18 to 37, participated in this experiment. To control the influence of language on reading as well as the impact of diversity in human vision on eye tracking, we pre-screened participants and recruited native English speakers who had normal to corrected-to normal vision. Upon completing the experiment session, each participant received $30.

4.2 Experiment Design

Experiment had a within-subject design. Each participant completed four search tasks in an experiment session that lasted up to 1.5 hours. The tasks were designed based on prior work [16] to be at two complexity levels: simple and complex, shown in Table 2. Two types of search user interfaces were designed based on a commercial test search engine developed by Search Technologies Corp[1][17]. To control the order effect, we created 32 rotations of task complexity levels and user interface types with a constraint that UI is switched after two tasks. A rotation was assigned to each participant in a random order. In each search task, participants were required to read task description, complete pre- and post- questionnaires, and search information on Wikipedia using either of the two user interfaces. Participants were asked to bookmark and save the documents they judged relevant. There were no time limits set for search tasks. Before completing their experiment session, participants were asked to fill out exit questionnaires. Data analysis presented in this paper does not include user interface factor.

5. Data Analysis and Results

5.1 Data Preprocessing

Table 1 shows descriptive statistics for queries, documents examined and task descriptions visited.

	Median	Mean	SD
Not useful queries	9	7.81	4.81
Useful queries	6	6.00	2.17
Not useful documents	6	7.69	6.69
Useful documents	10	11.00	4.70
Task descriptions	18	17.53	9.66

Table 1. Descriptive statistics for queries, documents examined, and task descriptions visited per participants.

Before performing data analysis, we removed bad fixation data (marked by Tobii as validity=4) and standardized all predictors: each predictor variable was divided by its standard deviation.

[1] http://www.searchtechnologies.com/

5.2 Principal Component Analysis (PCA)

We used principal component analysis **[18]** to explore a possible factor structure of variables representing the cost in information search, such as $W_{notUseful}$, $T_{notUseful}$, F_{SERP}, $D_{notUseful}$, $F_{taskDescription}$, *and* $T_{taskDescription}$. Prior to PCA, we performed Kaiser-Mayer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity to ensure our sample supports valid PCA [19]. The result of KMO was 0.64, with Bartlett's Test yielding 600.46 ($p < 0.001$). Both values indicate that our own sample could satisfy the requisite assumptions for proceeding with PCA.

Type	Task scenario
Simple 1	You love history and, in particular, you are interested in the Teutonic Order (Teutonic Knights). You have read about their period of power, and now you want to learn more about their decline. You want to find out: What year was the Order defeated in a famous battle? And you also want to find out which army (or armies) defeated the Order?
Simple 2	You recently attended an outdoor music festival and heard a band called Wolf Parade. You really enjoyed the band and want to purchase their latest album. What is the name of their latest (full- length) album? And you also want to know when this band resumed their work together?
Complex 1	A local water conversation group requests ideas to expand their efforts. Currently, they pick up debris from local waterways and try to raise awareness about water pollution. In an effort to help out, you volunteer for the group but also, you want to expand their efforts. What other forms of land use are impacting waterways? Which forms of land use have are the highest impact to the environment?
Complex 2	A debate is underway after an international logging and mining corporation submitted a bid to buy a local nature reserve. The city needs more jobs but many residents are upset because they find selling a nature reserve as short sighted. And many people actively use the nature reserve for recreation and educational field trips. In an effort to be balanced with support for the community and to be fair to economic development, you decide to investigate both sides further. What are the small and large scale impacts of logging and mining? What are some economic considerations for land preservation? What are your recommendations to the city if the corporation's bid is successful?

Table 2. Task Descriptions

To determine the number of components to keep, we used parallel analysis [20]. Specifically, a random dataset was generated with the same number of rows and variables as that in our sample data. Based on the generated dataset, we then created a correction matrix and computed eigenvalues. The results of parallel analysis are shown in Figure 1 with red line. According to the results, we kept only two components that appeared above the red line.

Given the number of the components determined in parallel analysis, we ran PCA gain with oblique (Promax) rotation and the number of factors fixed to 2. The factor loadings for manifest variables were shown in **Table 3**. The standardized loadings of these manifest variables (highlighted in **Table 3**) ranged from 0.70 to 0.98. The value of $\boldsymbol{h^2}$ indicates the final communality estimate: the proportion of variance accounted for by retained components. A value of $\boldsymbol{h^2}$ <0.40 indicates that an item is less strongly correlated with its corresponding components [21]. Based on this criterion, all of the six variables in our research were correlated with their corresponding components.

Analyzing the results of PCA, we named the two factors *exploratory* process and *validation* process (shown in **Table 3**). The exploratory process was reflected in 3 variables: total time taken to enter *not useful* queries ($T_{notUseful}$), number of words used in all *not useful* queries ($W_{notUseful}$), and number of fixations on all SERPs (F_{SERP}). The validation process was reflected in three other variables: number of fixations on visited task descriptions ($F_{taskDescription}$), Total time of visiting task descriptions ($T_{taskDescription}$), and number of fixations on visited *not useful* documents ($D_{notUseful}$).

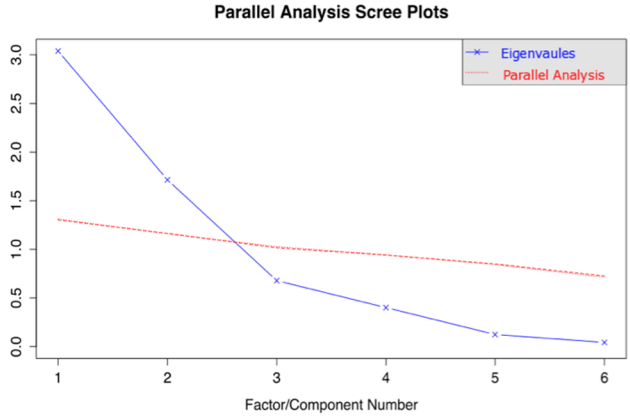

Figure 1. Revised scree plot showing parallel analysis results

In the exploratory process, people first define their goals and express their information problem by generating queries that are expected to retrieve the relevant materials. The results of executing different queries are then explored by examining search engine result pages (SERPs). In the validation process people confirm whether the collected information satisfies task goals [10] by examining documents and task descriptions.

Each of these two processes could span several stages of information search and that searchers can move back and forth between these processes. Roughly, the exploratory process takes place during Initiation, Selection and Exploration stages in Kuhlthau's ISP model; it corresponds to Define Problem, Select Source, Formulate and Execute Query stages in Marchionini's information seeking-process [9, 10]. The validation process takes place during Formulation and Collection stages in Kuhlthau's model, and it corresponds to Examine Results stage in Marchionini's information seeking-process.

	Exploratory Process	Validation Process	h^2
$T_{notUseful}$	**0.98**	-0.09	0.91
$W_{notUseful}$	**0.94**	-0.07	0.85
F_{SERP}	**0.75**	0.17	0.68
$F_{taskDescription}$	-0.08	**0.98**	0.92
$T_{taskDescription}$	0	**0.94**	0.88
$D_{notUseful}$	0.04	**0.7**	0.51

Table 3: Factor loadings for each manifest variable

6. DISCUSSION AND CONCLUSION

Using PCA, we attempted to identify a possible factor structure of variables that represented different costs in information search. Our results demonstrated that these variables load onto two very well separated factors. This enables us to suggest that these two factors correspond to the two different processes in information search. We observe that they correspond quite well to exploratory process and validation process.

To the best of our knowledge, this point has not been considered in previous studies. As mentioned before, in exploration process, people try to issue as many useful queries as they can to maximize the likelihood of finding information that satisfies their information needs; during the validation process people confirm whether useful information has been found. In particular, validation process involves high level cognitive activities, during which people need to be aware what information has been gained from information seeking and whether the information meets their information needs. The evaluation of the cost in validation process possibly enabled people to monitor and control when to modify or enter new queries so that they can lower their cost and maximize gain in search.

Our findings can inform further studies on the cost of information search. For instance, in the previous studies, the different kinds of costs were linearly combined. Based on our findings, in the future study, we can examine relationships between the costs coming from the exploratory process and validation process and ask for example, *are they linearly correlated with each other? Do they have different effect size on the information searching stopping behavior*?

Moreover, in our study, we have attempted to consider cognitive factor in the estimation of the cost in information search. For instance, we used the number of eye fixations to reflect people's cognitive efforts in reading. In the future study, we plan to examine how participants' working memory (WM) ability affects their performance and subjective perception of the cost of information search. It has been found that WM has a significant correlation with reading comprehension ability [22]. Thus individuals with low WM capacity are expected to exert more cognitive effort and suffer higher costs when examining documents, as compared to those with high WM capacity.

Limitations of our study include, firstly, PCA was used to find a possible underlying factor structure of a set of variables; it cannot be used to confirm the factor structure. Our study focused on the exploration of possible relations between the variables representing different costs in information search. In a future study, we intend to utilize confirmatory factor analysis (CFA) [20] to verify the factor structure that have been found in this paper. In addition, we made several simplifying assumptions about the *not useful* queries and documents. Lastly, owing to the limitation of our data sample, we had no additional dataset to the examine the gain of information search. We will try to address these limitations in our follow up studies.

7. ACKNOWLEDGMENTS

This research was supported, in part, by IMLS Career award to Jacek Gwizdka # RE-04-11-0062-11.

8. REFERENCES

[1] Smucker, M. D. and Clarke, C. L. A. 2011. Time-based calibration of effectiveness measures. In *Proceedings of SIGIR'2011*. ACM, New York, NY. 95-104.

[2] Hemmer, E. 2013. *Information Seeking Stopping Behavior in Online Scenarios*. Peter Lang Pub Inc.

[3] Speier, C. and Morris, M. G. 2003. The influence of query interface design on decision-making performance. *Mis Quart*, 27, 3, 397-423.

[4] Azzopardi, L. 2011. The economics in interactive information retrieval. In *Proceedings of SIGIR'2011*. ACM, New York, NY. 15-24.

[5] Azzopardi, L. 2014. Modelling interaction with economic models of search. In *Proceedings of SIGIR'2014*. ACM, New York, NY. 3-12.

[6] Azzopardi, L., Kelly, D. and Brennan, K. 2013. How query cost affects search behavior. In *Proceedings of SIGIR'2013*. ACM, New York, NY. 23-32.

[7] Varian, H. R. 1999. Economics and search. In *Proceedings of ACM SIGIR Forum*. ACM. 1-5.

[8] Stigler, G. J. 1961. The economics of information. *The journal of political economy*, 213-225.

[9] Kuhlthau, C. C. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*.

[10] Marchionini, G. 1997. *Information seeking in electronic environments*. Cambridge university press.

[11] Hauff, C., de Jong, F., Kelly, D. and Azzopardi, L. 2010. Query quality: user ratings and system predictions. In *Proceedings of SIGIR'2010*. ACM. 743-744.

[12] Kelly, D., Cushing, A., Dostert, M., Niu, X. and Gyllstrom, K. 2010. Effects of popularity and quality on the usage of query suggestions during information search. In *Proceedings of SIGCHI' 2010*. ACM, New York, NY. 45-54.

[13] Wu, W.-C., Kelly, D. and Huang, K. 2012. User evaluation of query quality. In *Proceedings of SIGIR'2012*. ACM, New York, NY. 215-224.

[14] Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J. and Zhang, X. M. 2011. Task and user effects on reading patterns in information search. *Interact Comput*, 23, 4, 346-362.

[15] Gwizdka, J. and Cole, M. 2010. Eye Movement Patterns and Interaction for High Level Information Seeking.

[16] Kelly, D., Arguello, J., Edwards, A. and Wu, W.-c. 2015. *Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework*. ACM, City.

[17] Zhang, Y. and Gwizdka, J. 2014. Effects of tasks at similar and different complexity levels. *Proceedings of the American Society for Information Science and Technology*, 51, 1, 1-4.

[18] Jolliffe, I. 2014. *Principal Component Analysis*. John Wiley & Sons, Ltd.

[19] Worthington, R. L. and Whittaker, T. A. 2006. Scale development research - A content analysis and recommendations for best practices. *Couns Psychol*, 34, 6, 806-838.

[20] Zhang, Y., Zhang, J., Lease, M. and Gwizdka, J. 2014. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of SIGIR'2014*. ACM, New York, NY. 435-444.

[21] Norris, M. and Lecavalier, L. 2010. Evaluating the Use of Exploratory Factor Analysis in Developmental Disability Psychological Research. *Journal of Autism and Developmental Disorders*, 40, 1, 8-20.

[22] Daneman, M. and Carpenter, P. A. 1980. Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19, 4, 450-466.