Modeling Decision Points in User Search Behavior

Paul Thomas
CSIRO,
Australia
paul.thomas@csiro.au

Peter Bailey
Microsoft,
Australia
pbailey@microsoft.com

Alistair Moffat
The University of Melbourne,
Australia
ammoffat@unimelb.edu.au

Falk Scholer
RMIT University,
Australia
falk.scholer@rmit.edu.au

ABSTRACT

Understanding and modeling user behavior is critical to designing search systems: it allows us to drive batch evaluations, predict how users would respond to changes in systems or interfaces, and suggest ideas for improvement. In this work we present a comprehensive model of the interactions between a searcher and a search engine, and the decisions users make in these interactions. The model is designed to deal only with observable phenomena. Based on data from a user study, we are therefore able to make initial estimates of the probabilities associated with various decision points.

More sophisticated estimates of these decision points could include probabilities conditioned on some amount of search activity state. In particular, we suggest that one important part of this state is the amount of utility a user is seeking, and how much of this they have collected so far. We propose an experiment to test this, and to elucidate other factors which influence user actions.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*.

Keywords

Retrieval experiment; evaluation; system measurement.

1. INTRODUCTION AND BACKGROUND

When a user selects a search engine to help solve their information need, a complex human-computer interaction transpires. The search engine responds to a user's query with a search results page, containing links to documents and other resources including query reformulation suggestions. The user may find sufficient information just in the search results page, or by clicking on linked documents, or may need to issue additional queries and repeat the process.

While the start of the information seeking task is clearly indicated to the search engine by the user (on first entering a query), there are no such direct equivalents to indicate the completion of the information seeking task, let alone whether the interaction was successful or not. The intermediate steps of the interaction process are open to instrumentation however. By capturing various micro-activities within the information seeking process, we can gain a better understanding of the conditional probabilities of each decision.

In this work, we make a number of contributions. First, we present our model of user–search engine behavior interaction relevance accumulation, and decisions. Our model, which clarifies and expands on a number of previous models, incorporates examination of result summaries within a search results page, examination of individual documents contingent on clicking on the link from their summary, query reformulation, and search engine switching. Second, we make use of behavior interaction data from an existing user study to quantify the corresponding averaged decision probabilities within our model. Finally, we outline a follow-up experiment which would enable interaction decisions to be examined in more detail, and support the construction of a predictive model of user behavior.

Related work Other researchers have proposed various models to characterize user–search engine interaction behavior. Building on the Expected Search Length ideas of Cooper [4], Dunlop [7] proposed number-to-view graphs that explored the number of documents a user wished to view versus the number they had to view to find them. Dunlop also characterized search engine interface and presentation aspects within the same framework, calling them time-to-view graphs. Other general models of search interaction include the Anomalous States of Knowledge framework proposed by Belkin [2], the Information Search Process 6-stage task-based framework by Kuhlthau [10], and the Information Foraging model of Pirolli and Card [13], inspired from anthropological theories of optimal food seeking strategies to describe information seeking behavior. Like the Information Foraging model, Azzopardi [1] recently proposed an economic theory of user behavior to describe the choices made during searching, hypothesizing that users seek to minimize cognitive load. Azzopardi's theory both explains user activity and predicts likely search behaviors based on properties of a search system, such as response speed.

Turpin et al. [17] examined how result summaries might impact evaluation measures, separating out the relevance of a summary from the relevance of an entire document; however, the probability of clicking a document was fixed based on its underlying relevance, which is likely an oversimplification of real user behavior. The Tolerance to Irrelevance concept articulated by de Vries et al. [6] investigated search behavior in the context of multimedia and XML retrieval systems. Here, users move to the next search result in a list

IIiX'14, Regensburg, Germany.

once their individual “tolerance to irrelevance" threshold has been reached within the content of a current result.

The Cascade model proposed by Craswell et al. [5] hypothesizes a cascading model of examination down a ranked list of results, where if the result summary appears relevant to the user, there is a corresponding probability of the user clicking on the document and thereby concluding their search. This model does not consider query reformulation or obtaining utility from summaries alone. The Rank-Biased Precision metric of Moffat and Zobel [11] hypothesizes a task- and user-dependent probability of persisting in examining documents in a ranked list (and vice versa, a probability of ceasing examination), where the probability is independent of what has already been examined and whether or not the document is relevant. In subsequent work, Moffat et al. [12] characterize relevance utility metrics in terms of adaptive user models, focusing on the probability of the user continuing to examine documents based on what has been seen in the ranking until that point. Smucker and Clarke [15] model relevance accumulation from a time-dependent aspect, arguing for a gain contingent on the amount of time allotted by a searcher.

Switching between different search engines has also been studied extensively, chiefly by White and fellow researchers, in the context of understanding the motivations for changing from one search engine to another, using large-scale log data to extract user behavior features for predicting probabilities for switching to occur (see White and Dumais [18] and Guo et al. [8] as key works in this area). Due to limitations of space, we do not discuss Markov model-based approaches to induction of user intent.

2. A MODEL FOR USER BEHAVIOR

Figure 1 is one possible representation of the general sequence of actions followed – either explicitly or implicitly – by a user as they search. We emphasize that this is a user model only in a limited sense: it only considers actions which are observable by a search engine. Clearly there are corresponding mental states, and decisions and actions which do not involve a search engine, but these are in principle unobservable without interrupting a user’s interactions, or making use of (currently distracting) experimental apparatus to map brain activity. We argue that unobservable decisions and actions, which do not induce any sort of change in interaction, are in and of themselves much less interesting to search engine operators – in this case it’s not the thought that counts.

The first decision of a searcher, prior to entering the process shown in Figure 1, is to use web search tools to address an information need, rather than, say, asking a colleague, or picking up a reference book. Once that decision is made, a search service is selected, and a first query formulated and entered; this is where we can first observe the user. Assuming a standard search results page, the user begins examining the result summaries that are presented, maintaining a notional location i in the ranking that is their current position, and varying that position upward and downward as they examine summaries [16]. We can think of this process as an initial inspection of summaries, seeking to establish if one or more offers markedly better prospects for information satisfaction than do others that are nearby.

After browsing some number of summaries in this way, the user chooses one of them for more careful inspection. At this point, it may be that the information need is satisfied out of the summary (that is, the summary relevance of the ith document, s_i, is 1); or it might be that the linked document needs to be inspected, and a determination made as to whether it is relevant (document relevance $r_i = 1$) or not ($r_i = 0$); or it might be that after focusing on the summary, the user decides that it wasn’t helpful after all ($s_i = 0$). If

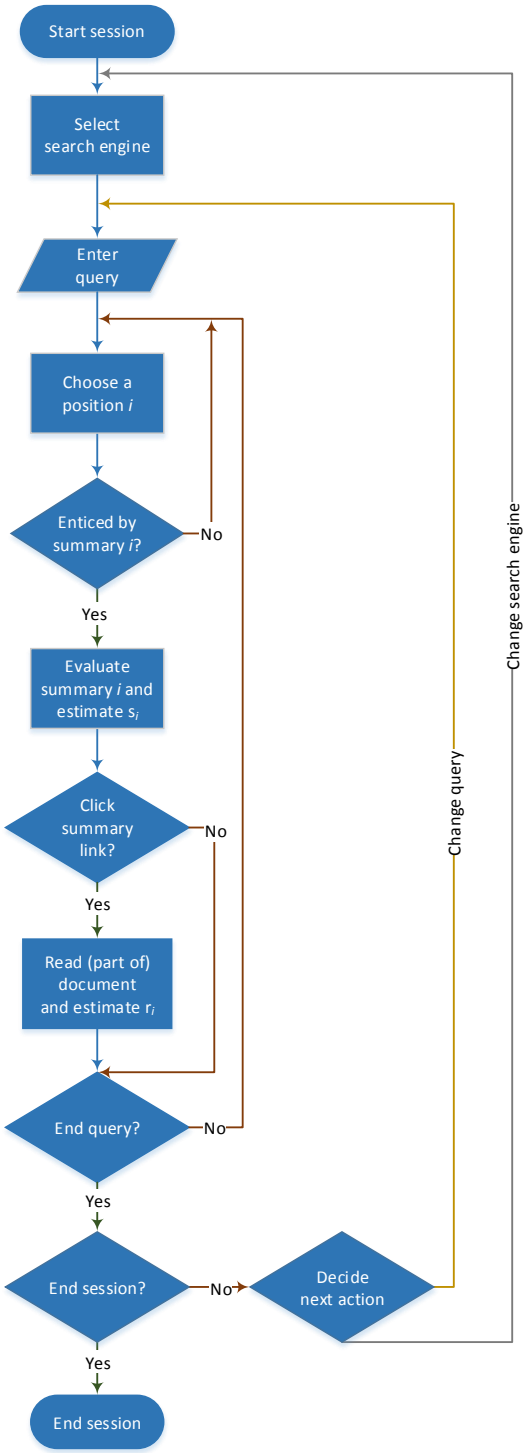

Figure 1: Decision points and processes for user search activity.

either $r_i = 1$ or $s_i = 1$ as a result of inspecting the ith summary, the user has made progress towards their information goal.

Such progress is then reflected in subsequent actions. In particular, the user may continue examining the search results list, changing the position i of the answer item that they choose to examine. Alternatively, the user can decide that they are unlikely to obtain sufficient further utility out of the set of results for this query to warrant further investment of effort, and end the query. They then select among

three alternative choices – they can end the session; or, if they don't wish to do that, can stay with the same search provider and reformulate their query; or, they can switch to a new search provider and execute a new initial query. Whichever decision they make, they will do so in the context of the information accumulated to date in their search, including a sense of how much progress they have made towards their information need; and how likely they anticipate it is that further progress can be made after spending more time.

Therefore, in addition to the observable actions, the user also evolves "state" as they follow the process mapped out in Figure 1, and this state informs the decisions made along the way. In particular, we posit that the user has a notion of how much *utility* they sought when they commenced their information seeking, a quantity that we refer to as T; and that they maintain a subconscious estimate corresponding to $\sum r_i + \sum s_i$, the amount of utility that they have accumulated so far through the course of the search. We argue that these quantities in particular condition the likelihood of decision outcomes. For example, if T is high (as is the case with an informational query) and the sum of relevance accumulated so far is low (as might happen with a "hard" topic, or a poor-quality search service), then reformulation will have a bigger relative probability. Similarly, if reformulation has already occurred multiple times, and accumulated relevance remains low, then there is likely to be an increased probability of taking the "change search engine" edge out of the "decide next action" decision point.

Another aspect of "state" is relevance, at both the summary and the document level. For example, in the event that a summary was not considered useful ($s_i = 0$), the decision made in "Click summary link?" will most probably be "No", whereby the user's implied assessment of the document's utility is also $r_i = 0$. While failure to find utility in a summary or document may assist the user in recalibrating their understanding of the information task (and therefore T), we argue that the amount of utility gained towards T by this negative outcome is still zero.

While we have presented Figure 1 as a complete flow chart, there are many more options in addition to the ones shown. Other transitions that could be added include those arising from: additional information resources shown on the page such as advertisements; distractions encountered through subsequent reading/browsing behavior; and interruptions leading to lack of continuity. The diagram is intended to illustrate the main high-level transitions that take place, rather than be exhaustive.

3. USER STUDY

To gauge the extent to which flows through Figure 1 can be quantified, we analyze the data associated with a previous user study [12]. In that study, a total of $n = 34$ subjects were asked to undertake a sequence of search tasks using an anonymized search interface, an instrumented browser that recorded all interactions together with timestamps, and with eye-tracking hardware monitoring their gaze locations. Users were provided with text descriptions of information needs; invited to formulate queries in order to address those needs; had their behavior monitored while they looked at the search engine results pages; and were explicitly asked, for each snippet that they clicked, whether the document they then viewed was useful in terms of helping to answer the original information need. Users were free to access second and subsequent pages for the same query, or to reformulate the query to obtain fresh result pages; and were also free to move on to the next task as soon as they felt that they had accumulated "enough" answer material to allow the information need to be dealt with. The only way in which the instrumented sessions diverge from the hypothesized arrangement shown in Figure 1 is that the subjects in our study were not able to switch search engine, the rightmost path in the diagram.

A total of six information need statements were provided, of a range of query types. For each participant, three (rotated) topics displayed results retrieved using the API of a commercial search service. Two traces were lost due to recording errors, but the remaining $34 \times 3 - 2 = 100$ interaction traces provide a rich record of user behavior, including time spent looking at and reading snippets; reading documents; and reformulating queries.

4. ANALYSIS

Based on the user study data, we derived information for each node in the diagram, representing the number of observations or the aggregate results of each decision. In particular, for this initial exploration we are interested in first-order estimates of the flow through each node: that is, how often are searchers enticed by summaries, how often do they click a link, how often do they end a query? The results are presented in Table 1.

Sessions and queries At the top end of the diagram, the number of *sessions* is fixed at 100. *Selecting a search engine* was not possible in the user study, so we cannot estimate the flow through this point. The number of *queries entered*, 208, was recorded by our search UI.

At the bottom end of the diagram, it is trivial to record the flow through the "yes" and "no" branches from *end query* (each query ends exactly once, so the "yes" count must be the same as the number of queries issued) and from *end session* (for which similar logic applies). Again, it was not possible from our setup to record flows on the *change search engine* branch.

Evaluations and clicks Flows through the middle part of the diagram, representing interactions following a single query, are harder to estimate. Gaze-tracking hardware allowed us to record each participant's fixations, or the points on screen which they looked at. Defining areas of interest corresponding to the 10 result summaries on each search results page enabled the frequency with which ranks were viewed to be determined, giving an estimate of the number of times that participants *chose a position i*.

In our model, users choose a position repeatedly until a summary is sufficiently "enticing" that it becomes a focus of attention and seems worth investing the effort to skim, read, or click on that summary. However, an "evaluation" in this sense need not involve reading: trust and rank bias (see Joachims et al. [9]) may mean the evaluation is as simple as recognizing where the summary lies on the page, for example, or the evaluation may involve more complex interactions dependent on features such as the number of query words in boldface. This makes "enticement" difficult to detect from trace data. For this exploration we used some simple heuristic rules:

- Any fixation on a summary, followed by a click on that summary, indicates enticement.
- Any sequence of fixations on a summary identifiable as "skimming" or "reading" behavior, according to the classifier of Buscher et al. [3], is considered to be evidence of enticement.
- Any sequence of fixations on a summary which this classifier marks as ambiguous, but which has characteristics of both reading and skimming, is taken to be evidence of enticement.
- Any sequence of fixations on a summary which looks like neither reading nor skimming, and where there is no click, is not considered to be evidence of enticement, and is more likely to simply be a brief eye movement.

Start session	100 sessions
Select search engine	not observed*
Enter query	208 queries
Choose a position i	2872 fixation sequences
Enticed by summary i?	
$\rightarrow$ yes	after 931 of 2872 sequences (32%)
$\rightarrow$ no	after 1941 of 2872 sequences (68%)
Evaluate summary i	931 evaluations*
Click link for summary i?	
$\rightarrow$ yes	after 301 of 931 examinations (32%)
$\rightarrow$ no	after 630 of 931 examinations (68%)
Read (part of) document	301 reads
End query?	
$\rightarrow$ yes	after 208 of 931 evaluations (22%)
$\rightarrow$ no	after 723 of 931 evaluations (78%)
End session?	
$\rightarrow$ yes	after 100 of 208 queries (48%)
$\rightarrow$ no	after 108 of 208 queries (52%)
Decide next action	
$\rightarrow$ change query	after 108 of 108 queries
$\rightarrow$ change search engine	not observed*

Table 1: Flows and first-order probabilities, based on 100 sessions with a competitive search engine. *: see text for further discussion.

Given these rules for counting enticement, we are able to derive the flow through either branch of the *enticed by summary i* decision.

Finally, *clicks on summary links* and the associated document reading activities were recorded directly by our search interface.

5. DISCUSSION AND FUTURE WORK

The numbers of observations at each point in the model are summarized in Table 1. For each decision point, the number of times each choice was observed is also listed. As just one instance of the data that has been gathered, only around 32% of the summaries that are observed are later clicked on.

This is a simple, context- and memory-free formulation but one which could be used as a crude simulation of a search user (reiterating that we are only interested in modeling observable interactions with a search engine). With appropriate relevance judgments, such a simulation could even be used to drive batch evaluations; this has something of the flavor of the simulations of user variance proposed by Smucker and Clarke [14], but with more detail.

If the structure presented in Figure 1 is plausible, then it would be useful to develop more nuanced, individual models of each decision point – models which take into account appropriate contextual features. This would enable the generation of synthetic interaction traces for evaluation, and in turn allow distinctions to be drawn between the many effectiveness metrics that have been proposed. It would also enable predictions of how changes to a search engine might influence user behavior to be made. Moffat et al. [12] have considered this for the choice of position i, and it is possible to similarly model the other decision points. Doing so would be plausible even with the limited data from the user study described previously – for example, the probability of clicking on a link could be estimated as a function of the time spent reading it, and the probability of enticement could be estimated as a function of T, the accumulated relevance. If we were to limit these sub-models to include only the most significant factors, the complexity would be tractable, and the model might have useful predictive or explanatory power.

A further study is then required to provide data to build more sophisticated models that involve more wide-ranging relationships and hence properly explore the paths through Figure 1. The decision to switch engines is well-studied [8, 18] and fairly well-understood. However, other decisions in our model are doubtless influenced by several factors to do with the user, their task, and their interaction history. While it is possible to imagine what these factors might be, it is also possible to guess wrong. In future work we plan to first carry out an initial study using a think-aloud protocol – perhaps supplemented by reviews of videos – to elicit the factors at play when users make each of the decisions from our model. From these, and informed by theories such as information foraging and the economic model of information interaction, it should be possible to formalize a set of factors and make particular predictions of which will be influential, and when. A second, larger study will then test these predictions, through direct observation or by manipulating the factors and looking for changes in observed behavior. The model that emerges will provide a reference point for future IR evaluation.

Acknowledgment This work was supported by the Australian Research Council's Discovery Projects Scheme (projects DP110101934 and DP140102655). We thank Dingyun Zhu for his help.

References

[1] L. Azzopardi. The economics in interactive information retrieval. In *Proc. SIGIR*, pages 15–24, 2011.

[2] N. J. Belkin. Anomalous states of knowledge as a basis for information-retrieval. *Canadian Journal of Information Science-Revue Canadienne Des Sciences De L Information*, 5:133–143, May 1980.

[3] G. Buscher, A. Dengel, and L. van Eist. Eye movements as implicit relevance feedback. In *Proc. CHI (Extended Abstracts)*, pages 2991–2996, 2008.

[4] W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation*, 19(1):30–41, 1968.

[5] N. Craswell, O. Zoeter, M. J. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proc. WSDM*, pages 87–94, 2008.

[6] A. P. de Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *Proc. RIAO*, pages 463–473, 2004.

[7] M. D. Dunlop. Time, relevance and interaction modelling for information retrieval. In *Proc. SIGIR*, pages 206–213, 1997.

[8] Q. Guo, R. W. White, Y. Zhang, B. Anderson, and S. T. Dumais. Why searchers switch: Understanding and predicting engine switching rationales. In *Proc. SIGIR*, pages 335–344, 2011.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. SIGIR*, pages 154–161, 2005.

[10] C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *JASIS*, 42(5):361–371, 1991.

[11] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Information Systems*, 27(1):2:1–2:27, 2008.

[12] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. CIKM*, pages 659–668, 2013.

[13] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106(4):643, 1999.

[14] M. D. Smucker and C. L. A. Clarke. Stochastic simulation of time-biased gain. In *Proc. CIKM*, pages 2040–2044, 2012.

[15] M. D. Smucker and C. L. A. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.

[16] P. Thomas, F. Scholer, and A. Moffat. What users do: The eyes have it. In *Proc. AIRS*, pages 416–427, 2013.

[17] A. Turpin, F. Scholer, K. Jarvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *Proc. SIGIR*, pages 508–515, 2009.

[18] R. W. White and S. T. Dumais. Characterizing and predicting search engine switching behavior. In *Proc. CIKM*, pages 87–96, 2009.