Personalised Search Time Prediction using Markov Chains

Vu Tran
University of Duisburg-Essen
vtran@is.inf.uni-due.de

David Maxwell
University of Glasgow
d.maxwell.1@research.gla.ac.uk

Norbert Fuhr
University of Duisburg-Essen
norbert.fuhr@uni-due.de

Leif Azzopardi
University of Strathclyde
leif.azzopardi@strath.ac.uk

ABSTRACT

For improving the effectiveness of *Interactive Information Retrieval (IIR)*, a system should minimise the search time by guiding the user appropriately. As a prerequisite, in any search situation, the system must be able to estimate the time the user will need for finding the next relevant document. In this paper, we show how Markov models derived from search logs can be used for predicting search times, and describe a method for evaluating these predictions. For personalising the predictions based upon a few user events observed, we devise appropriate parameter estimation methods. Our experimental results show that by observing users for only 100 seconds, the personalised predictions are already significantly better than global predictions.

1 INTRODUCTION

Interactive Information Retrieval (IIR) is a complex, non-trivial process where searchers undertake a variety of different actions over the course of a search session [7]. With a large number of variables that can impact upon how an individual searches, modelling the IIR process is extremely complex and has attracted a large amount of attention from the community (e.g. [1–4, 6, 10, 11, 14, 17]). For quantitative modelling of IIR, the *Interactive Probability Ranking Principle (IPRP)* [6] formulates a general principle for structuring the interaction between a user and a system. It assumes that the user performs a sequence of decisions about choices offered to him or her by said system. Each choice involves a certain degree of *effort* (or *cost*) for evaluating it, and when it is accepted (with some probability), it results in a certain *benefit*. The IPRP then derives a criterion for the optimum ordering of the choices such that the expected benefit of the decision list is maximised. As the IPRP is a rather general framework, it does not specify the type of costs and benefits to be considered.

A natural choice for measuring costs and benefits is to use time. The economic approach for modelling IIR [2] uses the same *'currency'*. It is straightforward to measure the cost of specific actions (e.g. the average time it takes a user to formulate a query, to look at a result snippet, or scan through a potentially relevant document). However, estimating benefit is a much more complex issue, as there is no simple method for doing this for the various actions possible in a specific situation (e.g. how much does it help reformulating the query or inspecting a results list item?). Tran and Fuhr [15] proposed regarding the (saved) *Time To the next Relevant document (TTR)* as benefit. However, they were only able to estimate TTR values retrospectively, and did not try to make any predictions.

We address in this paper the issue of TTR estimation as an important step towards estimating the benefit of potential user actions. This will allow us to apply the IPRP for *user guidance*. However, retrieval time depends heavily upon the specific user due to individual factors, such as typing and reading speed. Thus, general TTR estimates are of little help. Instead, we require a *personalisation* of these estimates. Moreover, time estimates are closely related to time-based *evaluation* [12] of IIR as shorter times yield improved quality in terms of time-based measures.

To the best of our knowledge, this is the first paper aiming at estimating search times. As a preliminary study, we will only regard a simplified version of the general problem: instead of estimating times for all possible actions and situations, we only look at the time from the first/next snippet (after the first query or a relevant document) to the next relevant document. While these estimates themselves may be of little practical value, the methods described here can be used as a baseline for further research focusing on situation- and action-specific estimates. To this end, we focus on the following two research questions.

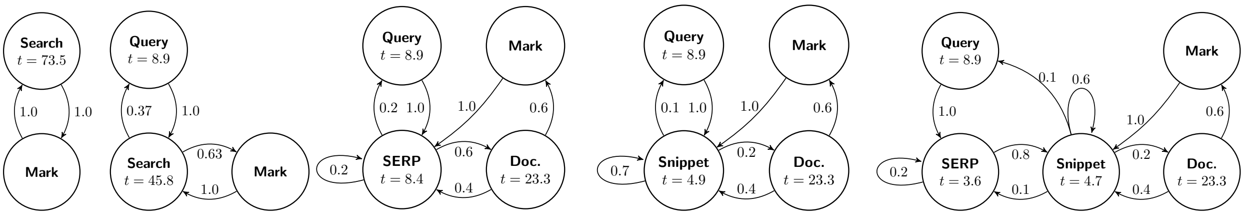
RQ1 Is it possible to attain reasonable TTR estimates, or do actual search times vary too much to make such predictions feasible?

RQ2 Can we personalise these estimates so that the average prediction error is smaller for these individual estimates?

2 RELATED WORK

Several different approaches have been proposed for modelling complex IIR processes. Zhang and Zhai [17] presented the *card model* as a theoretical framework for optimising the contents of the screen presented in a specific situation. As optimising criterion, they used information gain, which in terms of the IPRP can be regarded as a heuristic approach for estimating the difference between cost and benefit. However, it is unclear if and how information gain is related to evaluation criteria for entire search sessions.

In terms of general user modelling, Azzopardi [2] presented *Search Economic Theory (SET)*, based upon the approach of the IPRP framework [6]. With SET, user effort was measured via use of a



(a) MM1 **(b) MM2** **(c) MM3** **(d) MM4** **(e) MM5**

Figure 1: Five state diagrams – from *MM1* at subfigure *(a)* to *MM5* at subfigure *(e)* – representing the five Markov chains used for this study. Each of the diagrams highlights the states and transitions between them. Also included are the times and transition probabilities when each of the models were trained over *33 searchers, over two topics, yielding 66 sessions* (refer to Section 4 - *Predicting Interactions*), using the complete interaction data from each of their search sessions.

cost function. Using simulated interactions with cognitive load as the cost, Azzopardi [2] compared a variety of search strategies, examining the cost of interaction for a given level of expected output, or gain. Kashyap et al. [8] define a cost model for browsing facets to minimise the cost of interaction, and thereby increasing the usefulness of the interface.

These models commonly use cost (effort) and gain (benefit) measures to maximise the expected gain, although there are only few studies that actually estimated them. Tran and Fuhr [15] combined eyetracking data with system logs to model the search process as a Markov chain, where a searcher would transition between a variety of different states, including (re)formulating a query, examining the attractiveness of snippets, the examination of documents, and selecting relevant documents. With this Markov chain, they were able to estimate values for the IPRP with effort as the time spent on each state, and benefit saved as the TTR. The authors then extended the Markov chain to a more detailed one [16], where each result rank has its own state. By estimating the expected benefit for each state, they were able to tell the user at which rank it is better to formulate a query (instead of going further down the result list). Similar to this, Smucker and Clarke [13] modelled the *switching behaviour* of users engaging with ranked lists which provide different levels of gain and show at what point it is optimal to *'switch'*.

3 USER DATA AND MARKOV MODELS

For this study, we were provided with interaction logs from 48 subjects who participated in a user study, each using the same search system to undertake ad-hoc topic retrieval over the TREC AQUAINT collection [9]. Subjects undertook two time-limited search tasks, with each task limited to a total of 20 minutes (1200 seconds), and were assigned to one of four experimental conditions[1]. Over the two search tasks, subjects *on average* submitted 11.7 queries and examined 38.5 documents. In this preliminary analysis, we use a subset of the interaction data from 36 subjects which were assigned to the first three conditions. This is due to the fact that there were no significant differences between the first three conditions; the remaining 12 subjects differ significantly in terms of interaction times from the first three conditions.

Considering the interaction log data we acquired, we propose five different models based upon discrete time, discrete state Markov chains with costs as times spent on each state (refer to Figure 1). We start with a very simple model (***MM1***) and increase the complexity with each model (up to ***MM5***). The aim of this approach is to cover log files with different levels of granularity. As a baseline, we predict the average search time, which is represented here as Markov model ***MM1*** comprising the two states *(i) search* and *(ii) marking a document as relevant*. In the second model ***MM2***, we added state, *(iii) query*, for formulating a query. We added more details in the search process by replacing the search with *SERP*, examining the *Search Engine Results Page (SERP)* and *document*, assessing a document for relevance, naming this model ***MM3***. For ***MM4***, we changed SERP interactions to *snippet* interactions. Instead of simply modelling all the time spent on a SERP as a single state, we split it into one state per snippet examined. These simplistic representations of SERP/snippet interactions were then replaced by a fifth, amalgamated Markov chain, ***MM5***, where we consider both the SERP interaction time and snippet time. Here, SERP time is assumed as the time spent after submitting the query or asking for the next 10 results, until the requested SERP time is displayed (due to the underlying search engine, this took several seconds). The snippet time then refers to the actual time spent per snippet (subject to the approximations described below).

4 EXPERIMENTAL METHOD

Interpreting Log Data The user study log file contains a series of events: *query box focus*, *query submitted*, *view SERP page x*, *snippet hovers* (both in and out, with the mouse cursor), and *view* and *mark* documents. Each event has a timestamp, with document-centric events also containing the original rank. We considered the query state as the point from which a searcher focused on the query box to submitting their query. Examining a document was interpreted as the duration from which a document was displayed to a subject to the time that they either marked the document as relevant, or left the document altogether (i.e. returning back to the corresponding SERP). SERP time was considered as the duration from which SERP x was displayed to the subject, to the point that they left the SERP by either: focusing on the query box (to reformulate); viewing a document; or viewing the next/previous SERP y. For ***MM5***, the SERP time was considered as the duration from viewing SERP x to

[1] Space restrictions limit a more thorough explanation of the user study; refer to [9] for further details.

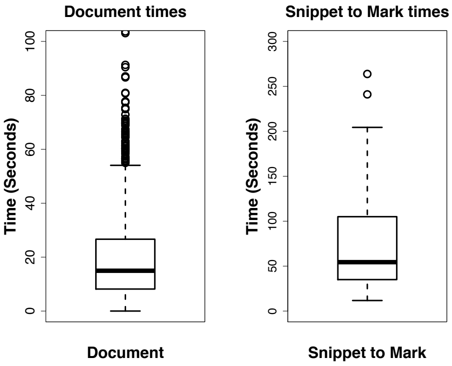


Figure 2: Time distribution on documents and snippet-to-mark from the actual user study log data (refer to Maxwell and Azzopardi [9]).

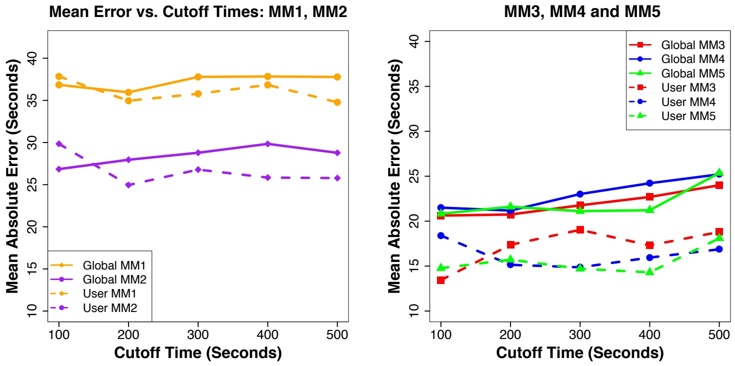


Figure 3: The mean absolute error of the predictions for each Markov chain model over the cutoff times (refer to Section 4). Note that the absolute error is compared over the various interaction data cutoff times.

the time the first result item was examined (via the first recorded hover in event), and from the previous action (e.g. marking a document) to the viewing of SERP y. Snippet time was considered as the duration the subject spent examining a snippet. Since hovering events proved to be unreliable, we had no direct information on these events. Instead, we assumed that the subject looked sequentially through the snippets, and when he or she clicks on a document, we divided the time since the SERP examination began by the rank of the viewed document. Based upon this assumption, we created the corresponding number of snippet events. In the case where no snippet on the SERP was clicked, we created artificial snippet events with the average duration per snippet derived from the observed clicks.

As can be seen from Figure 2, document times varied substantially, with a fairly large number of outliers (those that are more than 3.5 standard deviations away from the mean, i.e. above 58 seconds). As it is impossible to predict such outliers, one reasonable solution would be to discard these sessions. However, since we have only a limited amount of empirical data, we decided to keep these sessions, but to 'cap' the outlier document times, by assuming that the user did not spend more than 58 seconds per document.

Measures Examined The most obvious time to predict was the span from query formulation until finding the first relevant document. However, since users were asked to find as many relevant documents as possible – and with our limited number of observed search sessions – it was more sensible to be able to make predictions for each relevant document found. After finding a relevant document, users typically go back to a SERP and look at the next snippet. For this reason, the most appropriate time to be considered is the one from the first/next snippet viewed (or SERP in case of ***MM3***) to marking a document relevant.

Estimating Times and Probabilities The transition probability between any two states s_i and s_j is estimated using a maximum likelihood estimation: $P_{r(s_i,s_j)} = N_{ij}/N_i$, where N_{ij} is the number of times we observed a transition from state s_i to state s_j, and N_i is the total number of transitions from state s_i to any other state in the training data. In a similar way, the expected time spent for each state (*Query*, *SERP*, *Snippet*, *Document*) is computed as the average of the observed times in these states respectively.

With our Markov chains, we estimate the so-called *mean first passage time*, which is the expected time from one state to another. We explain this method for the case of ***MM4*** here. Let us denote the four Markov states q, s, d and m, the time in these states t_q, t_s, t_d and t_m, and the transition probability from state x to state y as p_{xy}. The expected times T_q, T_s and T_d for reaching the mark state from the query, snippet/SERP or document states respectively can then be computed via the following linear equation system.

$$\begin{aligned} T_q &= t_q + p_{qs}T_s \\ T_s &= t_s + p_{sq}T_q + p_{ss}T_s + p_{sd}T_d \\ T_d &= t_d + p_{ds}T_s \end{aligned}$$

We derived the actual observed behaviours from the user study log data. The actual time $\hat{T}_s$ (snippet or SERP to mark) was calculated as $\hat{T}_s = (\hat{T}_{lM} - \hat{T}_{fS})/|M|$. Here, $\hat{T}_{lM}$ is the timestamp of the last mark in the session. Since we are making predictions for the remainder of a session at specific cutoff times, $\hat{T}_{fS}$ is the timestamp of the first snippet seen for which we have not yet reached a marked document. Finally, $|M|$ denotes the number of documents marked in the remainder of the session.

Predicting Interactions We worked with 72 sessions (36 subjects with 2 topics each) in total. A pilot study showed no major differences between the two topics; as such, we consider both topics together. We used 12 fold cross-validation for all tests, meaning our training group and our test group contained 66 and 6 sessions respectively. When selecting subjects for training and testing, we created stratified samples by selecting the first three experimental conditions. This helped us to factor out the effects across the different experimental conditions. We evaluate our predictions with the actual observed data and present the mean absolute errors.

Global models are trained over 66 entire subsamples of session data as our baselines and tested on the remaining subsample. Sessions were cutoff into time slices, from 0 seconds (at the initial query focus) to periods ranging from 100 up to 500 seconds in steps of 100 seconds. These cutoffs provided us with five variations of the same log data, with each increase in time providing more interaction data. We then used the remainder of the sessions to evaluate our predictions by comparing the predictions from our generated models against the observed behaviours.

Personalised models are built from cutoff data of each individual subject. For building these models after some short observation time, we face the problem of parameter estimation: some transitions or states even may not yet have been observed for a specific subject. For the states, we use the following Bayesian formula to estimate the time: $T_x = \overline{T}_x v + Cm/v + m$, where $\overline{T}$ is the time of the global model at the given point of time, and v is the total number of observations until that point. C is the mean time of that state across the entire session, and m is the weight given to the prior estimate that is based on the distribution of average times derived from the entire session.

As for the probabilities for our personalised models, even a few observed events will not lead to good estimates using the standard maximum likelihood technique. Thus, we instead use Bayes' estimates with beta priors where the parameters of the beta distribution are derived from the overall distribution of probabilities in the training sample via the *method of moments* [5].

5 RESULTS

Figure 2 shows the overall distribution of the actual snippet times. Even after capping the document viewing times as described above, there is still a large variance in these times, making the task of predicting these times extremely difficult.

The mean absolute error of the various models investigated are depicted in Figure 3, where we show these errors for various cutoff times. All approaches consider the snippet-to-mark times for marks occurring after the cutoff time. For the user models, user-specific parameters are derived from the observations occurring before the cutoff time (i.e. these models are trained for some time, allowing them to make predictions for session's remainder). Significance tests are achieved using 2-tailed paired t-tests, with $p < 0.05$.

Given that the average snippet-to-mark time is 73.5 seconds, the relative errors are not very satisfactory for most of the approaches – which is at least partially due to the high variance of the values to be predicted. Comparing the performance of the five models ***MM1-5***, it is obvious that the first two models are outperformed by the three latter ones. Specifically, ***MM1*** produces very high errors due to the fact that it does not distinguish between querying and result examinations. With this distinction, both global and user-specific models of ***MM2*** perform significantly better than those of ***MM1***, showing that the complexity of a user's interaction requires a model with a certain level of detail. With even more details, ***MM3-5*** show much better performance than ***MM1-2***, although the improvement seems to stagnate when comparing ***MM3***, ***MM4*** and ***MM5*** to each other. This shows that increasing detail boosts the results only to a certain point, and after this point, results increase moderately.

Comparing the global models with the user-specific ones, we can see that the latter models are much better, even with very little training time. Only for the simple, poor-performing models ***MM1*** and ***MM2***, personalisation is of limited value. For ***MM3-5***, after only 100 seconds of training, all user-specific models are significantly better than their corresponding global ones, and this holds true also for the rest of the session. Earlier results without Bayes' estimators showed a different picture: the personalised models were worse than the global ones until 400 seconds.

6 CONCLUSIONS AND FUTURE WORK

User guidance for maximising the expected benefit of a search session is a major goal of quantitative models of IIR. In this paper, we devised a method for estimating these benefits in terms of search time, which is directly related to time-based evaluation measures. Moreover, we have shown that we can significantly improve global estimates by generating user-specific predictions after having observed the user for a short time.

Although the models regarded here are still fairly simple, these results are rather promising. Future work will focus on more complex models, considering (for example) rank positions of snippets, or the number of query reformulations. Only with these extensions will it be possible to guide the user (e.g. *go to the next rank*, or *reformulate your query* [16]). Moreover, we have considered only one type of search task here. Models for other types of tasks will also have to be developed (as well as classification methods for recognising the current user's task type). Nevertheless, the work presented in this paper is an important first step along this path.

Acknowledgements This work was supported by *ESF ELIAS* grant nos. `7109` (Tran) and `7271` (Maxwell), *GSF* grant no. `FU 205/26-1` (Tran), and *EPSRC* grant no. `1367507` (Maxwell). Our thanks to the three anonymous reviewers for their feedback.

REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. 2011. Find it if you can: a game for modeling different types of web search success using interaction data. In *Proc. 34th ACM SIGIR.* 345–354.

[2] L. Azzopardi. 2011. The Economics in Interactive Information Retrieval. In *Proc. 34th ACM SIGIR.* 15–24.

[3] F. Baskaya, H. Keskustalo, and K. Järvelin. 2013. Modeling behavioral factors in interactive information retrieval. In *Proc. 22nd ACM CIKM.* 2297–2302.

[4] A. Borisov, I. Markov, M. de Rijke, and P. Serdyukov. 2016. A Context-aware Time Model for Web Search. In *Proc. 39th ACM SIGIR.* 205–214.

[5] K.O. Bowman and L.R. Shenton. 2007. The beta distribution, moment method, Karl Pearson and RA Fisher. *Far East J. of Theoretical Statistics* 23, 2 (2007), 133.

[6] N. Fuhr. 2008. A Probability Ranking Principle for Interactive Information Retrieval. *Information Retrieval* 11, 3 (2008), 251–265.

[7] P. Ingwersen and K. Järvelin. 2005. *The Turn: Integration of Information Seeking and Retrieval in Context.*

[8] A. Kashyap, V. Hristidis, and M. Petropoulos. 2010. FACeTOR: Cost-driven Exploration of Faceted Query Results. In *Proc. 19th ACM CIKM.* 719–728.

[9] D. Maxwell and L. Azzopardi. 2014. Stuck in Traffic: How Temporal Delays Affect Search Behaviour. In *Proc. 5th ACM IIiX.* 155–164.

[10] D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. 2015. Searching and Stopping: An Analysis of Stopping Rules and Strategies. In *Proc. 24th ACM CIKM.* 313–322.

[11] P. Pirolli and S.K. Card. 1999. Information foraging. *Psychological Review* 106 (1999), 643–675. Issue 4.

[12] M.D. Smucker and C.L.A. Clarke. 2012. Time-based Calibration of Effectiveness Measures. In *Proc. 35th ACM SIGIR.* 95–104.

[13] M.D. Smucker and C.L.A. Clarke. 2016. Modeling Optimal Switching Behavior. In *Proc. 1st ACM CHIIR.* 317–320.

[14] P. Thomas, A. Moffat, P. Bailey, and F. Scholer. 2014. Modeling Decision Points in User Search Behavior. In *Proc. 5th ACM IIiX.* 239–242.

[15] V. Tran and N. Fuhr. 2012. Using Eye-Tracking with Dynamic Areas of Interest for Analyzing Interactive Information Retrieval. In *Proc. 35th ACM SIGIR.* 1165–1166.

[16] V. Tran and N. Fuhr. 2013. Markov Modeling for User Interaction in Retrieval. In *MUBE SIGIR Workshop.*

[17] Y. Zhang and C. Zhai. 2015. IR As Card Playing: A Formal Model for Optimizing Interactive Retrieval Interface. In *Proc. 38th ACM SIGIR.* 685–694.