Towards Detecting Wikipedia Task Contexts

Hanna Knaeusl
Chair for Information Science
University Regensburg
Germany
hanna.knaeusl@ur.de

David Elsweiler
Chair for Information Science
University Regensburg
Germany
david.elsweiler@ur.de

Bernd Ludwig
Chair for Information Science
University Regensburg
Germany
bernd.ludwig@ur.de

ABSTRACT

Wikipedia is a resource used by many people for many different purposes. We posit that it might be beneficial to alter the content or the way content is presented depending on the task context. Here we describe a small pilot lab study to investigate features of interaction that might help to infer the contextual situation surrounding wikipedia search tasks. We describe our effort to collect data and analyse relationships between the features and the assigned task context.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Preference Elicitation, Info Seeking Behaviour

Keywords

Eyetracking, Wikipedia

1. INTRODUCTION

Information portals such as Wikipedia represent rich sources of information covering an incredibly broad range of topics. Many Wikipedia entries are also long and can cover aspects ranging from overviews and introductions to more detailed descriptions of advanced aspects that are perhaps only suitable for topic experts. Single pages can also contain not only text, but images, info-graphics, lists and navigational information. Previous research suggests that these resources will have several different contexts of use. For example, Marchionini [11] identifies three main types of search tasks, all of which are applicable to Wikipedia: *Lookup* tasks include finding answers to specific questions, known-item searches or navigating to specific pages. These tasks are contrasted with *exploratory search* tasks, which include *learn* tasks, where the aim is to acquire larger amounts of knowledge and achieve an enhanced understanding of a given topic, and *investigate* tasks, where the user makes use of found information and continues to contribute to or generate knowledge in some way. Elsweiler et al. [4] provide an additional task dimension, distinguishing between work-oriented tasks where information is required to complete some job and casual-leisure tasks, where the aim is more pleasure-focused, e.g. to pass time, to relax, to be entertained etc.

Wikipedia contributors are encouraged to create pages in a way that meets the needs of as many users as possible by including information on a topic with sufficient quantity, quality and completeness and structuring the content in a way that makes sense generally. Nevertheless, one could imagine that different content or different presentations of the same content might be more suitable in specific contexts. For example, lookup tasks may be best supported when facts in an article are presented as a list that can be scanned easily. In such scenarios, content such as images may be less helpful and perhaps even distracting. Contrastingly, in casual-leisure situations, users may want to focus on multimedia content or have information presented in a way that encourages browsing and information discovery.

We believe examples like this suggest there may be benefit in moving away from static pages, which try to cater for all usage situations, to dynamic pages that are generated appropriately based on the context of use. As a first step towards exploring this hypothesis, in this paper, we investigate how the context of use – the task type being performed – might be detected automatically from user-interactions with the system. We want to establish if the way the user interacts with the system, e.g. his mouse and keyboard interactions, eye movements, and click behaviour can provide implicit feedback regarding the usage scenario and user goals.

With this aim in mind, we present a small pilot study that allows us to evaluate a methodology for detecting the features of interaction that might help us infer the contextual situation surrounding a user's search task. We collect interaction data in the context of a controlled laboratory study and analyse relationships between the features of interaction and the assigned task context. The data show that for the small number of users in our study, the behaviour exhibited when completing tasks of different types is very different; users interact with different types of content in different ways. Further, we provide evidence that it is possible, at least for some users, to predict these behaviours based purely on mouse and keyboard interactions.

2. RELATED WORK

In the IR community a large amount of work has been performed to establish if interaction data can be used as a surrogate for explicit relevance judgements. This is known as implicit relevance feedback. Early research in this area demonstrated a correlation between the time spent reading a

action	label	description
Read	RE	User is reading text
Scan	SC	User scans content e.g. headlines, lists or whole page
Examine	EX	User examines element
Navigate	NV	User navigates

element	label	element	label
Headline	HD	Text passage	TX
List	LI	Introduction	IN
Picture	PI	Info Box	IB
Charts, tables etc.	IG	Links in Wikipedia	WI
Other navigation	ON		

Figure 1: Annotation labels for the user actions during Wikipedia search and for the gazed elements

document and explicit relevance judgements [12]. Although this has been disputed in naturalistic situations [10], White and Kelly show that when task type is taken into account clear signals can be found[16]. Other studies have shown that the amount of scrolling on a Web page [3], click-through for documents in a browser [9], bookmarking behaviour [7] and eye movements during the search [2]can all be used as implicit feedback to improve retrieval performance.

Interaction data can also be used as a means to predict user emotions. For example, Fox et al., show that query log features can be used to predict searcher satisfaction [6] and Feild et al.[5] used interaction data and physical sensors to predict levels of user frustration with high accuracy.

A third group of studies show correlations between different styles of interactions e.g. for some users visual attention on the screen can be predicted via mouse coordinates [15] We believe that the interaction style, the emotional state of the user and the motivating task context will be intrinsically related and that the work done previously suggests it may be possible to predict the task based on interaction data. We explore this in a small pilot study below.

3. DATA COLLECTION

In this section we provide details of the data collected and explain the motivation behind recording the data.

3.1 Study Design

Data was collected via a laboratory based user study with 4 users. The participants were information science students (1 male, 3 female) aged between 20 and 30. All of the participants were experienced wikipedia users and were comfortable using the wikipedia search facilities. Although this user population is not large or diverse enough to provide generalisable results, it is sufficient for our aims, which were to evaluate and improve the methodology and get a sense for the feasibility of our ideas.

Each participant performed 6 Wikipedia search tasks (2 of each of the 3 types of interest - lookup, learn and casual-leisure). The tasks were presented in the form of a simulated scenario and were ordered randomly to minimise learning effects. Example tasks for each type are shown in Figure 2.

After initially greeting the participant, the experimental procedure was explained in person. Then, to prevent biases, the participant was led automatically through the experiment on screen, with task descriptions, questionnaires and a web-browser window appearing when appropriate. The experimenters observed the tasks remotely in an adjoining room, where the participant's screen was mirrored.

lookup

action	TX	ON	PI	IN	IB	IG	WI	LI	HD
EX	0	0	0	0	0	42	0	0	0
NV	0	0	0	0	0	0	46	0	0
RE	0	0	0	23	0	23	0	27	0
SC	53	0	0	24	18	0	0	59	12

learn

action	TX	ON	PI	IN	IB	IG	WI	LI	HD
EX	0	0	89	0	0	93	0	0	0
NV	0	2	0	0	0	0	52	0	0
RE	1872	0	0	72	0	0	0	93	0
SC	172	0	6	2	0	0	0	62	285

casual-leisure

action	TX	ON	PI	IN	IB	IG	WI	LI	HD
EX	0	0	137	0	2	85	0	0	0
NV	0	11	0	0	0	0	105	0	0
RE	1876	0	6	274	1	0	0	90	32
SC	177	0	2	8	6	0	0	60	134

Table 1: Absolute frequencies of content elements for actions for the investigated task types

3.2 Data Collected

We collected a large amount of data from each participant before, during and after the study.

Questionnaires: A pre-study questionnaire collected demographics, search experience, and experience with wikipedia of the participants. Pre-and post-task questionnaires elicited perceptions of the task and domain knowledge, of success and the experience including emotional aspects, and finally a post-study questionnaire provided general impressions of the experiment.

Eyetracking Data: We recorded participant gaze patterns using an SMI RED eye-tracker. The associated BeGaze software recorded videos files of screen interactions with an additional layer indicating the area of the screen where the user is focusing his gaze. We manually annotated these complete overlaid video sequences with two labels. The first describes what the user is doing ("action"). This is a simple coding scheme but aligns with reading psychology research [14, 13]. It was the annotator who decided which action to code at what moment by following the focus displayed in the layer on top of the recorded screen. The second label describes the content ("element") being focused on and is derived from the elements available in Wikipedia pages. The label was assigned when the focussed on an area on the screen so long that the annotator could assume the element in the area was perceived.The full set of labels for actions and elements is presented in Fig. 1. The intuition behind the labels was that the style of reading for different task types and the content elements used will be very different. By labelling videos in this way we could test this intuition empirically.

Browser Logs: We instrumented the firefox web-browser to log all user interactions during the search process.

Timestamp information was used to align interaction data from different sensors.

Lookup: Last night you watched a documentary about the sinking of the Titanic. Suddenly you wonder how many passengers were on board when the catastrophe happened. Search in Wikipedia for this information.

Learn: Friends from abroad are visiting Germany and you plan to travel together to visit the small but beautiful city of Regensburg. As preparation for the trip you want to know more about the city and its history. Use Wikipedia to do this.

Casual-leisure: You have a few minutes before your class starts but you are already sitting in the lecture hall. Kill this time using wikipedia using the next six minutes to look at whatever topic(s) take your fancy.

Figure 2: Examples of the kinds of tasks assigned to study participants.

4. EVALUATION OF THE DATA

We analyse the data in two stages. First, in Section 4.1, we examine the distribution of video labels for different types of task to determine if users behave differently or focus their attention on different kinds of topics when completing different task types. Second, in Section 4.2, we show how these labels can, in turn, be predicted using interaction data from the eyetracker and browser. The first stage provides evidence that the user's preferences for content elements depends on the search task, endorsing our suggestion to customise web pages at run time. The second stage provides some evidence for our hypothesis that the interactions a user performs in a browser may be used to predict which actions he trying to complete and which content elements he is preferring at that moment.

	LO vs RE		**LO vs CA**		**RE vs CA**	
action	χ^2	p-value	χ^2	p-value	χ^2	p-value
EX	9	0.011	9	0.029	18	0.006
NV	9	0.011	9	0.011	18	0.001
RE	13	0.043	6	0.301	27	0.079
SC	36.563	0.064	45	0.039	45	0.039

Table 2: χ^2-tests for Different Distributions of Content Elements per Task Type (LO: lookup, RE: learn, CA: casual-leisure)

4.1 Reading Style and Content for Task-types

Technical difficulties meant we were only able to work with data for 6 casual-leisure, 4 lookup tasks and 4 learn tasks. We first divided the data into 500ms frames, allowing us to normalise the counts by task length, and counted relative frequencies of frames for which label combinations occur for each task type (see Table 1). Visually inspecting the distribution of content for actions, suggests the reading style and the elements of content interacted with were very different in different task contexts. This is confirmed by pair-wise comparisons using chi-squared tests for the distributions content elements for each possible pair of task types (see Table 2).

Examining the results in Table 2, we observe that all but one combination of action type shows highly significant differences in the distribution of content elements examined. The exception is the distribution of elements for lookup and casual-leisure tasks, which initially seems counterintuitive, as one would expect these two tasks to be very different. Below we summarise the main similarities and differences between the task-types and attempt to explain what these mean in the context of our work.

When completing lookup tasks, the participants do not typically read content, the exception being page introductions. Instead they scan large portions of the page very quickly, looking for the snippets of information that will satisfy their specific information need. They tend to scan a number of different kinds of content elements during tasks. This can be seen from Table 1 with counts being spread over text passages, introduction, info boxes, lists and headers. Images are noticeably missing from lookup tasks. It seems as if the participants have decided that for the tasks assigned, images will not useful and are able to avoid them.

Learn and casual-leisure differ from lookup tasks in that they both tend to be longer in time and have more interactions. They also both involve reading actions, which were rare for lookup. By this we mean that the user focuses attention on whole passages of text and attends the text from left to right and line by line. Another similarity between learn and casual-leisure tasks is the way that text passages are consumed, with the counts for these tasks being very similar. There are differences between learn and casual-leisure tasks, particularly in terms of the elements used other than text passages. During learn tasks the focus tended to be on headers, while for casual leisure, the focus was on elements such as introductions and info boxes, which allow the user to gain an overview of what a page is about and allow them to judge whether it is interesting or not. We assume that headers are useful for learn tasks because here there is a concrete information need i.e. users do not just need to find something that is interesting or not, but need specific informational content. In this sense headers will help the user determine whether a paragraph is worth reading or not.

4.2 Predicting Style and Content Preferences

To determine if the manually assigned labels can be predicted from interaction data alone, we calculated statistics for counts of the synchronous occurrences of video labels and input events for the 500ms frames introduced above. As we were searching for the simplest features possible (so they could eventually be computed easily during a browser session at runtime) we used the frequencies of the most common mouse events and the average saccade distance (i.e. eye movement) per frame as features. More precisely, for each frame we descretised these features into two levels: *low* and *high* based on the mean value over all frames.

Table 3 (left) gives an example for the information we computed from the raw log data. In order to understand whether the knowledge of the `mousemove` frequency is relevant for predicting user actions and content elements, we performed a series of χ^2-squared tests for all six search tasks for one of the test participants chosen at random (in total about 30 minutes of interaction). The results are reported in Table 3(right). With the exception of the rare `click` events, all features are highly significant. We interpret this as a positive indication that for individual users – depending on their personal interaction style (see [1, 8]) – it is feasible that the reading behaviour label could be predicted during a brows-

action	mousemove high	mousemove low	element	mousemove high	mousemove low
NV	5	6	IN	30	12
RE	18	5	IB	8	10
SC	41	18	WI	5	6
			LI	21	1

Task	scroll action	scroll el.	click act.	click el.	mousemove act.	mousemove el.	avg.sacc.dist act.	avg.sacc.dist el.
1	***	***	***	***	***	***	***	**
2	***	***				*	***	***
3	*	**		*	***	***	* **	
4	*	***	*			**		***
5	***	***			***	***	*	
6	***	***	**		***	***	**	***

Table 3: Frequency counts of user actions and `mousemove` events and of content elements and `mousemove` events occurring simultaneously (left). The table on the right shows the significance results for χ^2-squared tests.

ing session. The results of the χ^2-squared tests indicate that knowing at run-time whether the observed input events occur below or above average at any point of time increases the accuracy of predicting the video labels as annotated for that moment as the distribution $P(\text{action}|event = \text{low})$ differs significantly from the distribution $P(\text{action}|event = \text{high})$ for any annotated action and for any annotated element type. This oberservation opens the way for runtime prediction of the user action and preferred elements. From that information, the system can predict the current task type and use this information for generating content dynamically.

5. CONCLUSIONS

The preliminary data analysis we have presented provides clues that, firstly, reading behaviour and preferences for content elements depend on the surrounding task context and, secondly, both behaviour and preferences may be predicted for individual users based on their interaction style.

There are several limitations to this work. That we only have data from four participants from a relatively homogenous group means we cannot generalise. However, we claim that the presented methodology is well suited to address our long term research questions outlined in the introduction and the pilot has provided us with insight into how to improve a full study. In addition to resolving several technical challenges, we have learned that the great care will need to be taken when simulating tasks. For example, were few images looked at in lookup tasks, simply because of the tasks we chose? We also plan to look at more complicated prediction features and account for the fact that individual differences in participants (cognitive, reading style [14]) will exists and that users interact in different ways (people who follow eye movements with their mouse, people who don't) [15]. At EuroHCIR, we look forward to engaging with the broader HCI and IR communities to discuss the ideas in this paper; we are particularly eager to receive feedback on the next steps along this research path, including brainstorming solutions to some of the empirical design challenges of running such experiments and identifying and dealing with the many factors which should be incorporated in the full study.

6. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of SIGIR*, SIGIR '06, pages 3–10, 2006.

[2] G. Buscher, A. Dengel, and L. Van Elst. Eye movements as implicit relevance feedback. In *CHI'08: Extended Abstracts on Human Factors in Computing Systems*, page 2991–2996, 2008.

[3] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the IUI*, page 33–40, 2001.

[4] D. Elsweiler, M. L. Wilson, and B. Kirkegaard Lunn. *New Directions in Information Behaviour*, chapter Understanding Casual-leisure Information Behaviour. Emerald Publishing, 2011.

[5] H. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *Proc of SIGIR 2010,*, 2010.

[6] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inform. Syst.*, 23(2):147–168, 2005.

[7] Q. Guo and E. Agichtein. Ready to buy or just browsing?: detecting web searcher goals from interaction data. In *Proceedings of SIGIR*, pages 130–137, 2010.

[8] J. Huang, R. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. In *Proceedings of CHI*, CHI '12, pages 1341–1350, New York, NY, USA, 2012. ACM.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinki, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inform. Syst.*, 25(2), 2007.

[10] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of SIGIR*, page 408–409, 2001.

[11] G. Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, 2006.

[12] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieva. In *Proceedings of SIGIR*, pages 272—281, 1994.

[13] J. Nielsen. *Designing Web Usability*. New Riders, Berkeley, Calif., 2006.

[14] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psych. Bull*, 124(3):372—422, 1998.

[15] K. Rodden and X. Fu. Exploring how mouse movements relate to eye movements on web search results pages. In *SIGIR Workshop on Web Information Seeking and Interaction*, pages 29–32, 2007.

[16] R. W. White and D. Kelly. A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of CIKM 2006*, page 297–306, 2006.