Influence of Timeline and Named-entity Components on User Engagement

Yashar Moshfeghi[1⋆], Michael Matthews[2], Roi Blanco[2], Joemon M. Jose[1]
Yashar.Moshfeghi@glasgow.ac.uk, mikemat@yahoo-inc.com,
roi@yahoo-inc.com, and Joemon.Jose@glasgow.ac.uk

[1] School of Computing Science, University of Glasgow, Glasgow, UK⋆⋆
[2] Yahoo! Labs, Barcelona, Spain

Abstract. Nowadays, successful applications are those which contain features that captivate and engage users. Using an interactive news retrieval system as a use case, in this paper we study the effect of timeline and named-entity components on user engagement. This is in contrast with previous studies where the importance of these components were studied from a retrieval effectiveness point of view. Our experimental results show significant improvements in user engagement when named-entity and timeline components were installed. Further, we investigate if we can predict user-centred metrics through user's interaction with the system. Results show that we can successfully learn a model that predicts all dimensions of user engagement and whether users will like the system or not. These findings might steer systems that apply a more personalised user experience, tailored to the user's preferences.

1 Introduction

User engagement refers to the positive aspects of a users' interaction experience, in particular users' captivation by a technology [1]. Given the ubiquity of the choices on the web and the competitiveness of the market, applications nowadays are designed to not only be efficient, effective, or satisfying but also engaging [2]. Thus, a new vein of research is to identify system features that steer user engagement [3], which has become a key concept in designing user-centred web applications [1]. There has been great attention on retrieving named entities[3] [4, 5], and using the time dimension for retrieval [6]. Those approaches are evaluated exclusively focusing on a Cranfield-style paradigm, with little or no attention on user input, context and interaction. However, it is difficult to correlate user engagement with traditional retrieval metrics such as MAP [7]. This problem becomes exacerbated when the user has to cope with content-rich user interfaces that include different sources of evidence and information nuggets of a different

⋆ Work performed while intern at Yahoo! Research.

⋆⋆ **Acknowledgement.** This work was supported partially by the EU LiMoSINe project (288024).

[3] Named-entities are chunks of text that represent a real-world entity and which can be classified into a broad set of categories, such as person, date, organisation etc.

nature. This work studies the interplay between user engagement and retrieval of named-entities and time, in an interactive search scenario.

Further, we investigate if an automatic method can predict user-centred metrics, using users' interaction with the system and their demographics and search habits as an input. Given the increase of information-rich user experiences in the search realm [8], we leverage the amount of logged interaction data. Prediction of user preferences for web search results based on user interaction with the system has been studied previously [9]. In this work, we try to predict user-centred metrics of an IIR system *rather than* user preferences for its search results. Our positive findings could steer research into building search applications in which the layout and elements displayed adapt to the needs of the user or context.

To provide a use case for our investigation, we experiment with a news search system, which encourages interaction due to the information overload problem associated with the news domain. One way to facilitate user interaction in such scenarios is to develop new methods of accessing such electronic resources. For this purpose, we carefully varied the components of a news retrieval system page. We experimented with a timeline and named-entity component (enriched) or hiding them (baseline), while keeping everything else fixed, and tested whether adding these components can help improve user engagement. To study the predictability of the user centred metrics, we repeat our interactive experiments at two different points in time, with a tightly controlled setting. As an outcome of those experiments, we conclude that the user centred metrics can be predicted with high accuracy given their interaction with the system and their demographics and search habits are provided as an input.

This paper has two novel contributions: (i) the study of the effect of named entities and time in user-centred metrics such as user engagement, in an interactive search scenario, using a crowdsourcing platform; where through crowdsourcing, questionnaires and log data is collected for linking qualitative to quantitative user engagement metric; (ii) the study of predictability of the user-centred metrics given a number of features derived from the participants' demographics, search habits and interaction with the system (i.e. log data).

2 Related Work

User Engagement: User engagement is a multi-faceted concept associated with the emotional, cognitive and behavioural connection of user with a technological resource at any point of interaction period [1]. Through user engagement we understand "how and why people develop a relationship with technology and integrate it into their lives." [1] O'Brien and Toms [3] defined a model characterising the key indicative dimensions of user engagement: focused attention, aesthetics, perceived usability, endurability, novelty, involvement. These factors elaborate the user engagement notion over the emotional, cognitive and behavioural aspects. Subjective and objective measures are proposed to evaluate user engagement [1], the former being considered to be the most for evaluation. We use the subjective measures proposed by O'Brien et al. [3]. Objective

measures include subjective perception of time (SPT) and information retrieval metrics among others. SPT is calculated by asking participants to estimate the time taken to complete their searching task, which is compared with the actual time [1]. Interactive IR metrics are directly related to measuring engagement [1], and take into account users and their contexts. IIR evaluation is based on the idea of simulated search scenarios, where a subject is asked to follow a search scenario that specifies what, why, and in which context the user is searching. In this paper we follow the IIR evaluation framework.

Time and Entity Retrieval: Adding a time dimension to IR applications has gained increased attention of late with examples such as news summaries [10]. Alonso et al. [6] suggested the feasibility of automatically creating timelines from temporal information extracted from documents. Along this line, Koen et al. [11] augmented news articles by extracting time information, whereas, Ringel et al. [12] placed search results in a timeline for desktop search.

In addition to time, searching for named-entities is a common user activity on the Web, particularly in the news domain [5]. For example, there has been much work in entity search where the goal is returning the entities, such as people and locations, that are most relevant to a query [4]. The increased effort in developing entity search techniques and in building evaluation benchmarks [4] indicates its importance. In this work we combine timeline and named-entity features within an interactive news system to study their effects on user engagement using an IIR evaluation framework.

Why Crowdsourcing?: IIR evaluation has been used widely in IR where the experiments were conducted in laboratory-based environment and participants were introduced to a simulated search tasks controlled by researchers. Several drawbacks of this approach have been discussed recently, including: lack of generality due to their population bias and small sample size and lack of representativeness of real search scenarios due the artificial search environment [13]. Given these limitations, the popularity of using a crowdsourcing platform for performing user-based evaluation has increased rapidly. For example, some works have focused on capturing relevance assessments or labels for documents using crowdsourcing [14], whereas others have captured user interactions with search engines [15]. Some previous work was geared towards better understanding the nature and characteristics of these platforms [16] and their comparability with laboratory-based user study experiments [17]. In this paper we *do not* try to compare crowdsourcing platform with a laboratory-based environment, and solely use it to conduct our user study.

3 News System Description

User Interface: The user interface is built around four main components: the query-input component, the search result component, the timeline component and the named-entity component. The query-input component (shown in Figure 1 (A)) simply consists of a text box and search button, while the remaining components are detailed below.

Search Result Component: The core of the user interface is a search result component, which shows the retrieved articles upon submission of a query. Each article's information is presented using a snippet component (shown in Figure 1 (B)). In this component, in addition to the standard highlighted snippet text, there are lists of both the most important *keywords* and dates associated with the document. They serve to better summarise the document and also to provide additional clues for refining the search. An image is displayed if one is crawled with the news item and associated with the document which provides additional relevance clues. In the case of multiple images for a news item, we assigned the image top-ranked by a model learnt to find the most interesting picture, using features such as the image quality, presence or absence of faces, photo size, etc. For the purpose of the user study, clicking on an article title brought up a consistent modal screen containing the article contents which guaranteed that the user stayed within the news system and also ensured that the user experience was not affected by the differences in presentation by different news providers.

Timeline Component: Figure 1 (C) displays the timeline component. The component is split between two bands - the bottom band (trend graph), shows how the frequency of documents containing ***republican debates*** changes over the 6 months covered by the collection while the top band (topic timeline) uses a widget to display the titles of the top ranked articles. In our study, after submitting a query, the top 50 retrieved articles are placed on the topic timeline indicated by circles, to enhance and ease the interaction with them. Similar to the *Search Results* component, clicking on an article title on the topic timeline brings up the same modal screen containing the article contents. Such interaction is transparent to every other component since it does not refresh the page.

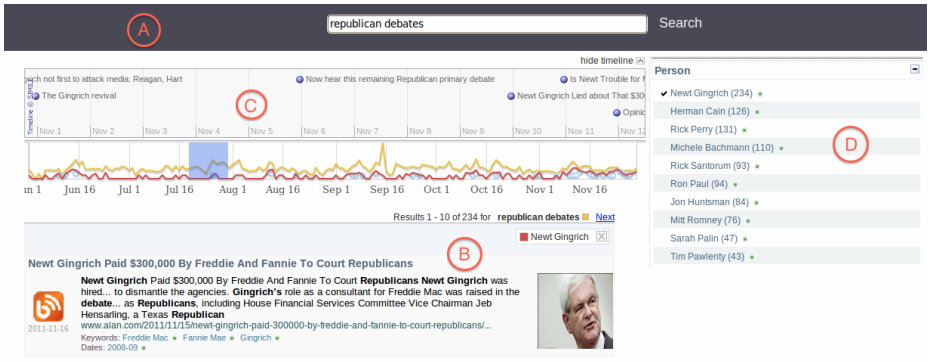

Fig. 1: Snapshot of the user interface, the main components have been identified: (A) query-input, (B) search result, (C) Timeline, and (D) named-entity component

Named-entity Component: Figure 1 (D) demonstrates an entity list panel. The entity list is modified as the query is refined allowing the user to easily

see how the important entities change for a given context. Clicking on an entity filters the ranked list so that only articles containing this entity are shown. It also updates the timeline with visual clues as to which entities have been selected by providing a trend line for the entity on the trend graph.

Baseline vs. Enriched System: For experimental purposes, the system was configured so that the timeline and entity components could easily be hidden from the view. The user interface of the baseline system consists of the search result components (Figure 1 (B)) along with a common query-input component (A). The user interface of the enriched system, in addition to these components, has two additional components: timeline (C) and named-entities (D).

User Tracking and Logging: All user actions were monitored and logged by the system including their queries, clicks, the overall length of time users spent on the system, and the length of time spent reading articles. Additionally, given that the mouse movements are correlated with user gaze [18], the system captured mouse events to determine the amount of time the user spent with the mouse over the main components of the user interface – the snippet section, the entity panel and the timeline topic and trend lines.

Corpus: The news system is built on a set of 820 different sites and 715 blog RSS feeds.[4] The RSS feeds were crawled nightly adding roughly 3000 blog entries, 3000 news entries and over 2000 related images. The news sources have topical and geographic diversity allowing for the selection of a variety of task topics and their presentation in a variety of formats. The data was gathered for a period of 6 months from June to November 2011 with the final collection consisting of around 1,000,000 text documents and 400,000 images[5].

Backend Processing: The news retrieval system is developed using a subset of analysis tools.[6] The resulting analysis is used to extract, from each document, all person, location and organisation entities and all time expressions that can be resolved to a specific day, month or year. The time expressions extracted are both explicit as in "September 2013" and relative as in "next month". Relative dates are resolved based on the *publication date* of the article and all dates are associated as *event dates* with the corresponding documents. To represent the most important concepts contained in the document, the top 10 tf-idf ranked entities in a document are assigned to it as *keywords* if they have occurred at least twice.

4 Experimental Methodology

We aim to answer the following research question: *can timeline and named-entity components improve user engagement in the context of a news retrieval system?*

[4] The blog RSS feeds are based on a list provided by the FP7 SYNC3 project `http://www.sync3.eu/`

[5] The crawl is available upon request from the Internet Memory Foundation (`contact@internetmemory.org`).

[6] The analysis tools include OpenNLP (`http://opennlp.sourceforge.net/`) (for tokenization, sentence splitting and part-of-speech tagging, and shallow parsing), the SuperSense tagger (`http://sourceforge.net/projects/supersensetag/`) (for named-entity recognition) and Time ML (`http://www.timeml.org/site/index.html`) (for annotating document with temporal expressions).

In the remainder of the paper, we devise a use case scenario for evaluating an IIR system, in particular we test whether our previously described enriched news search system can enhance user engagement.

A 'within-subjects' design was used in this study. The independent variable was the system (with two levels: *baseline*, *enriched*), which was controlled by the viewing timeline and named-entity components (enriched) or hiding them (baseline). The dependent variables were: (i) user engagement (involvement, novelty, endurability, usability, aesthetics, attention), and (ii) system preference.

Task: The search task was presented using a simulated information need situation. We introduced a short cover story that helped us describe to our participants the source of their information need, the environment of the situation and the problem to be solved. This facilitated a better understanding of the search objective and, in addition, introduced a layer of realism, while preserving well-defined relevance criteria. The simulated task was defined as follow: *"Imagine you are reading today's news events and one of them is very important or interesting to you, and you want to learn more. Find as much relevant news information as possible so that you can construct an overall (big) picture of the event and also cover the important parts of it."* The search task was presented twice to each participant with different search topics (we refer to as First and Second Task). We prepared a number of search topics that covered a variety of contexts, from entertainment and sport to crime and political issues, in order to capture participants' interests as best as possible (shown in Table 1).

Table 1: The topics for the simulated search task scenario.

First Task	**Second Task**
Thai Floods	Turkey's Earthquake
EU Crisis (Debt)	US Jobs/Unemployment
Occupy Wall Street	Libya (Gaddafi)
Baseball (World Series)	Basketball Strike
Michael Jackson Trial	Amanda Knox Trial

Mechanical Turk: We make use of Amazon's Mechanical Turk (M-Turk), as our crowdsourcing platform. It provides a convenient participant pool to draw upon to carry out many tasks, from simple labelling to more complex tasks related to opinions. The benefit would be reduced monetary cost and ease of engaging a large number of users in the study. The downside is potentially low quality data and in turn, the challenge is to improve and assure data quality. Much research has been done in the past to present techniques and settings which can be applied by the requesters to minimise spammers, multiple account workers, and/or those who put unacceptable amount of effort in their assignments and/or being able to detect them at a later stage of the process [19, 16]. As in [19], particular attention was paid in our experimental design to help motivate participants to respond honestly to the self-report questions and take the tasks seriously. For example, we have employed the multiple response technique

for our questionnaire which is known to be very effective and cost efficient to improve the data quality [19]. Browser cookies were used to guard against multiple account workers, and to avoid spammers, participants drawn from the M-Turk population were screened automatically based on location (United States) and HIT approval rate greater than 95%, as recommended in the literature [19, 16]. To reduce attrition, demographic questions were put at the beginning of the experimental procedure [19].

Procedure: Participants were instructed that the experiment would take approximately 60 minutes to complete, though they would be given 120 minutes between the time they accepted and submitted the HIT assignment. They were informed that they could only participate in this study once and they would not be paid if they had participated in any of the previous pilot studies. Payment for study completion was $5. Given the findings of Mason and Watts, we expect the increase in wage just to change the rate of incoming workers to accept the HITS, and not affect their performance [20]. The total cost of the evaluation was $510 including the cost of the pilot studies and some of the rejected participants, which we consider to be cost-effective. Each participant had to complete two search tasks, one for each level of independent variable (i.e. baseline and enriched system). The order in which each participant was introduced to the systems was randomised to soften any bias, e.g. the effect of task and/or fatigue.

At the beginning of the experiment (before accepting the HIT) the participants were given an information page explaining its conditions. They could only accept the HIT if they agreed with a consent form. Subsequently, participants were assigned to one of two systems (baseline or enriched) by clicking the link to the external survey. Next, they were given the *entry questionnaire* to fill in. Before beginning each task, participants read the task information followed by a *pre-search questionnaire*. The session was preceded with a brief training video, designed for the system (e.g. baseline or enriched), highlighting the most important user interface features using an example task. Each user was required[7] to watch the entire video before starting the search session, ensuring that each participant had the same level of knowledge of the system and its features.

In each task, users were handed five topics and asked to proceed with the one they found most interesting. For each topic, the subjects were given 10 minutes, during which they had to locate as many relevant documents as possible. Afterwards, they were redirected to the news system website. After completing the task, participants were redirected back to their survey to respond to the *post-search questionnaire*. Questions in the *post-search questionnaire* were randomised to avoid the effect of fatigue. At the end of the experiment the *exit questionnaire* was given to the participants and they were redirected to M-Turk to submit their HIT for completion.

Questionnaires: At the beginning of the experiment, the participants were introduced to an *entry questionnaire*, which gathered background and demographic information, and inquired about previous experience with online news, in particular, browsing and search habits to estimate their familiarity with news

[7] Users had to enter a completion token displayed at the end of the video to continue the survey.

retrieval systems and their related tasks. At the beginning of each task, the participants completed a *pre-search questionnaire*, to understand the reason why a particular topic was selected. At the end of each task, the participants completed a *post-search questionnaire*, to elicit subject's viewpoint on all user engagement dimensions. Finally, an *exit questionnaire* was introduced at the end of the study. In this questionnaire we gathered information about the user study in general: which system and task they preferred and why and their general comments.

Qualitative and quantitative measures: User engagement was measured considering six dimensions introduced by O'Brien et al. [3]: focused attention, aesthetics, perceived usability, endurability, novelty, and involvement. The different dimensions were measured through a number of forced-choice type questions. For example, involvement was measured by adapting three questions from [3]: (1) *I was really drawn into my news search task.* (2) *I felt involved in this news search task.* (3) *This news search experience was fun.* Participants were instructed to respond to each item on a 5-point scale (strong disagree to strong agree): *"Based on this news retrieval experience, please indicate whether you agree or disagree with each statement"*. In total, in each post-search questionnaire we have asked 31 questions related to user engagement (adapted from [3]), and randomised its assignment to participants. In order to quantitatively assess the impact of time and entity dimensions, we used number of clicks and submitted queries and total time spent to complete the tasks captured via monitoring the participant's interaction with the system.

Pilot Studies: Prior to running the actual user study, we run three pilot studies using 10 participants. In each iteration, a number of changes were made to the system based on feedback from the pilot study. For example, for each dimension we computed Cronbach's alpha[8] to evaluate the reliability of the questions adopted for each dimension. We finalised the questions of each dimension by confirming their Cronbach's alpha value (> 0.8). Other changes consisted of modifications to the questionnaires to clarify questions, modifications to the system to improve logging capabilities and improvements to the training video. After the final pilot, it was determined that the participants were able to complete the user study without problems and that the system was correctly logging the interaction data.

5 Results and Discussion

63 out of 92 users successfully completed the study. There was a relatively even split by condition, with 47% in the scenario where participants first used the baseline and then the enriched systems, and 53% conversely. We removed the incomplete surveys and eliminated participants who were found to have repeated the study after either abandoning it part-way through or had completed it once before. Finally, those participants who had completed the survey incorrectly were

[8] Cronbach's alpha is used as a measure of the internal consistency of a psychometric test score for a sample of subjects.

identified and eliminated, based on the conditions explained in the task description: (1) they had to visit at least three relevant documents for a given topic,[9] and (2) the issued queries should be related to the selected topic. We followed the design guidelines presented in [15] for the quality control in crowdsourcing-based studies, such as identifying suspect attempts by checking the extremely short task durations and comments that are repeated verbatim across multiple open-ended questions. As a result of this process, we present the experimental findings of our study, based on 126 search sessions that were successfully carried out by 63 participants. The 63 participants (female=46%, male=54%, prefer not to say=0%) were mainly under the age of 41 (84%) with the largest group between the ages of 24-29 (33.3%). Participants had a high school diploma or equivalent (11.11%), associates degree (15.87%), graduate degree (11.11%), bachelor (31.7%) or some college degree (30.15%). They were primarily employed for a company or organisation (39.68%), though there were a number of self-employed (22.22%), students (11.11%), and not employed (26.98%).

Main Results: Figures 2 shows the box plot for the user engagement analysis, for the two systems (baseline and enriched), based on the post-study questionnaire. The box plot reports, over the data gathered from 63 participants, five important pieces of information namely: the minimum, first, second (median), third, and maximum quartiles. We performed a paired Wilcoxon Mann-Whitney test between measures obtained for the enriched system for each user to check the significance of the difference with the baseline system. We use (*) and (**) to denote the fact that a dimension had results different from that of the baseline with the confidence levels ($p < 0.05$) and ($p < 0.01$) respectively.

As shown in Figure 2, the enriched system has a better median and/or mean and lower variance than the baseline system across all dimensions. This shows that substantial user engagement improvements can be achieved by integrating time and entity information into the system. The findings also show that participants are significantly more engaged both from cognition (considering endurability and involvement) and emotion (considering the aesthetics and novelty) aspects when time and entity dimensions of the information space are provided (i.e. enriched system). We did not find any statistically significant difference between the two systems for PST metric (with mean and standard deviation of 10.03, $\pm$ 5.22, and 10.12, $\pm$4.95, for the baseline and enriched system respectively). In addition, the exit questionnaire posed the question *"Please select the system you preferred? (answer: 1: First System, 2: Second System)"* (we refer to as System Preference), and overall, 76% of the participants preferred the enriched system better than the baseline system.

[9] To ensure the availability of relevant documents, two evaluators manually calculated the precision@1, 5, and 10 for all the topics and a set of queries issued by the participants. Precision@1, 5 and 10 were 0.85, 0.84, and 0.86 respectively, and judges had a very high inter-annotator agreement (Kappa > 0.9). This indicates that the queries the users issued into the system had good coverage and the ranking was accurate enough. This is further explained by the fact that the topics were timely and most news providers including in the index contained articles related to them.

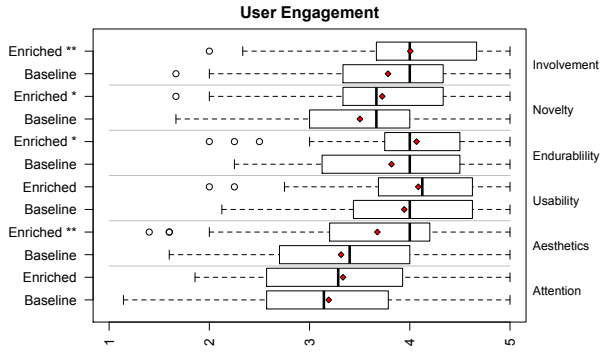

Fig. 2: Box plot of the user engagement based on the information gathered from 63 participants questionnaire. The higher the value, the higher the level of user agreement. The diamond represents the mean value.

Prediction of User-centred Metrics: We investigate whether user engagement and in a more general sense user-centred metrics can be predicted, given the participants' demographic and search habits information, and/or their interaction with the system. For this purpose, we used participants' age, gender, education, and occupation as the demographic features, the number of years they have used web search and online news systems, the frequency they engaged in different news search intention such as browsing, navigating, searching, etc. and the news domain they are interested in as the search habits features, derived from the entry questionnaire. Further, the total time they spent on each component and to complete a task, the number of clicks, retrieved documents, queries, and times they used the previous/next button, and other functionality of the systems as the interaction features, derived from log information. We chose the System Preference question and all the user engagement dimensions taken from exit and post-search questionnaire respectively. For System Preference question, we have a binary class of "-1" indicating the participant did not prefer the enriched system and "$+1$" otherwise. For the user engagement dimensions, we used the final value calculated by aggregating all the questions related to each dimension (presented in Figure 2). We transformed the values for each dimension to binary by mapping 4-5 to "$+1$" and otherwise to "-1" (similar transformation approaches have been used in the past [21]).

We learned a model to discriminate between the two classes using SVMs trained with a polynomial kernel, which, based on our analysis (not presented here due to the space limits), in the majority of cases, outperformed other SVM kernels (linear, polynomial, and radial-basis). We also tried other models such as bayesian logistic regression and decision trees but they underperformed with respect to SVMs. Table 2 shows the classification performance averaged over the

Table 2: The accuracy of the prediction for all the dimensions of the user engagement metric plus *system preference* question. The dimensions are presented in the columns in the following order: Involvement (INV), Novelty (NOV), Endurability (END), Usability (USE), Aesthetics (AES), Attention (ATT), and Preference(PREF). The features are presented in the rows with the following order: All the features (All), Demographics and Searching Habits (Demographics & Habits), Interaction History (Log). The best performing feature set for each dimension is highlighted in bold.

	User Engagement						***System***
	INV	NOV	END	USE	AES	ATT	PREF
Demographics & Habits	86.2%	77.2%	72.3%	89.2%	77.1%	**88.2%**	43.1%
Log	67.2%	71.1%	64.7%	67%	74.9%	72.5%	**87.4%**
All	**86.8%**	**81.1%**	**74.6%**	**89.2%**	**83.8%**	86.9%	50.6%

63 participants of the study[10] using 10-fold cross validation. Results indicate that for all the user engagement dimensions (excluding focused attention), the combination of all features leads to the best prediction accuracy. Remarkably, the machine learned model is able to predict with a low error all of the user and system metrics. Regarding the system preference question, user-system interaction features determine with high accuracy the participants' preference of a system (over 87%). Given these positive findings, it is possible to move towards personalised search applications in which the layout and elements displayed adapt to the needs of the user or context which in turn results in increasing the users' engagement as well as their preference of the system.

6 Conclusions

Given the competitiveness of the market on the web, applications nowadays are designed to be both efficient and engaging. Thus, a new vein of research is to identify system features that steer user engagement. This work studies the interplay between user engagement and retrieval of named-entities and time, in an interactive search scenario. We devised an experimental setup that exposed our participants on two news systems, one with a timeline and named-entity components and one without. Two search tasks were performed by the participants and through questionnaires, user engagement was analysed. Overall findings based on user questionnaires, show that substantial user engagement improvements can be achieved by integrating time and entity information into the system. Further analysis of the results show that the majority of the participants preferred the enriched system over the baseline system. We also investigated the hypothesis that user-centred metrics can be predicted in an IIR scenario given the participants' demographics and search habits, and/or interaction with the system. The results obtained across all the user engagement dimensions as well as System Preference question, supported our hypothesis. As future work, we will continue to study how user interactions can be leveraged to predict satisfaction measures and possibly build interfaces that adapt based on user interaction patterns.

[10] All the questions used in this study for user engagement as well as the participants' demographics, search habits and interaction data are available upon request.

References

1. Attfield, S., Kazai, G., Lalmas, M., Piwowarski, B.: Towards a science of user engagement (Position Paper). In WSDM Workshop on User Modelling for Web Applications (2011)
2. Overbeeke, K., Djajadiningrat, T., Hummels, C., Wensveen, S., Prens, J.: Lets make things engaging. Funology (2005) 7–17
3. O'Brien, H.L., Toms, E.G.: The development and evaluation of a survey to measure user engagement. J. Am. Soc. Inf. Sci. **61**(1) (2010) 50–69
4. Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., Attardi, G.: Ranking very many typed entities on wikipedia. In: CIKM, ACM (2007) 1015–1018
5. Demartini, G., Missen, M.M.S., Blanco, R., Zaragoza, H.: Taer: time-aware entity retrieval-exploiting the past to find relevant entities in news articles. In: CIKM. (2010) 1517–1520
6. Alonso, O., Gertz, M., Baeza-Yates, R.: Clustering and exploring search results using timeline constructions. In: CIKM. (2009) 97–106
7. Järvelin, K.: Explaining user performance in information retrieval: Challenges to ir evaluation. In: ICTIR, Springer-Verlag (2009) 289–296
8. Haas, K., Mika, P., Tarjan, P., Blanco, R.: Enhanced results for web search. In: SIGIR. (2011) 725–734
9. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: SIGIR. (2005) 154–161
10. Allan, J., Gupta, R., Khandelwal, V.: Temporal summaries of new topics. In: SIGIR. (2001) 10–18
11. Koen, D., Bender, W.: Time frames: Temporal augmentation of the news. IBM systems journal **39**(3.4) (2000) 597–616
12. Ringel, M., Cutrell, E., Dumais, S., Horvitz, E.: Milestones in time: The value of landmarks in retrieving information from personal stores. In: Proc. Interact. Volume 2003. (2003) 184–191
13. Kelly, D.: Methods for evaluating interactive information retrieval systems with users. Found. Trends Inf. Retr. **3** (2009) 1–224
14. Blanco, R., Halpin, H., Herzig, D.M., Mika, P., Pound, J., Thompson, H.S., Duc, T.T.: Repeatable and reliable search system evaluation using crowdsourcing. In: SIGIR. (2011) 923–932
15. Kittur, A., Chi, E., Suh, B.: Crowdsourcing user studies with mechanical turk. In: SIGCHI. (2008) 453–456
16. Kazai, G.: In search of quality in crowdsourcing for search engine evaluation. In: ECIR. (2011) 165–176
17. Zuccon, G., Leelanupab, T., Whiting, S., Jose, J., Azzopardi, L.: Crowdsourcing interactions-a proposal for capturing user interactions through crowdsourcing. In: CSDM at WSDM. (2011) 35–38
18. Guo, Q., Agichtein, E.: Towards predicting web searcher gaze position from mouse movements. In: CHI Extended Abstracts. (2010) 3601–3606
19. Mason, W., Suri, S.: Conducting behavioral research on Amazon's Mechanical Turk. Behavior research methods (June 2011) 1–23
20. Mason, W., Watts, D.J.: Financial incentives and the "performance of crowds". In: HCOMP. (2009) 77–85
21. Moshfeghi, Y., Piwowarski, B., Jose, J.M.: Handling data sparsity in collaborative filtering using emotion and semantic based features. In: SIGIR. (2011) 625–634