Comparative analysis of relevance feedback methods based on two user studies

Stephen Akuma, Rahat Iqbal, Chrisina Jayne, and Faiyaz Doctor

Original citation & hyperlink:
Akuma, S., Iqbal, R., Jayne, C. and Doctor, F., 2016. Comparative analysis of relevance feedback methods based on two user studies. *Computers in Human Behavior*, *60*, pp.138-146.
https://dx.doi.org/10.1016/j.chb.2016.02.064

DOI 10.1016/j.chb.2016.02.064
ISSN 1932-6203

Publisher: Elsevier

Computers in Human Behavior 60 (2016) 138–146

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Comparative analysis of relevance feedback methods based on two user studies

Stephen Akuma [a, *], Rahat Iqbal [a], Chrisina Jayne [b], Faiyaz Doctor [a]

[a] Department of Computing, Coventry University, Coventry, UK
[b] School of Computing Science and Digital Media, Robert Gordon University, Aberdeen, UK

ARTICLE INFO

Article history:
Received 15 June 2015
Received in revised form
12 February 2016
Accepted 15 February 2016
Available online 27 February 2016

Keywords:
Implicit feedback
User interest
Explicit feedback
Implicit indicators
Explicit rating
Recommender system

ABSTRACT

Rigorous analysis of user interest in web documents is essential for the development of recommender systems. This paper investigates the relationship between the implicit parameters and user explicit rating during their search and reading tasks. The objective of this paper is therefore three-fold: firstly, the paper identifies the implicit parameters which are statistically correlated with the user explicit rating through user study 1. These parameters are used to develop a predictive model which can be used to represent users' perceived relevance of documents. Secondly, it investigates the reliability and validity of the predictive model by comparing it with eye gaze during a reading task through user study 2. Our findings suggest that there is no significant difference between the predictive model based on implicit indicators and eye gaze within the context examined. Thirdly, we measured the consistency of user explicit rating in both studies and found significant consistency in user explicit rating of document relevance and interest level which further validates the predictive model. We envisage that the results presented in this paper can help to develop recommender and personalised systems for recommending documents to users based on their previous interaction with the system.

1. Introduction

As the amount of information available through the internet and various intranets continues to grow, effective retrieval of relevant information has become a challenging task. The tremendous amount of available information leads to information overload (Alhindi, Kruschwitz, Fox, & Albakour, 2015). This phenomenon has led to research on personalised information retrieval and recommender systems. Most of the current information retrieval systems are generic and do not provide task-specific information to users (Grzywaczewski & Iqbal, 2012; Jawaheer, Weller, & Kostkova, 2014). Hence, research in the area of personalisation has been found to be useful for addressing the problem of information overload by providing the users with relevant web documents based on their interest and current activity. This can enhance user search experience and improve efficiency. This can be achieved through relevance feedback based approaches. The process of gathering useful information about users in order to give them feedback or recommend documents based on their previous interaction with the system is called relevance feedback (Zemirli, 2012). Relevance feedback is affected by contextual factors like task type (Kellar, Watters, Duffy, & Shepherd, 2004; Li & Belkin, 2008). Tasks are activities that people attempt to accomplish to meet a particular goal. Identifying an accurate task stream is difficult because the demarcation of boundaries is not clear enough. However, common streams have classified different task along features like fact-finding vs information gathering (Kellar, Watters, & Shepherd, 2007). Task type help to infer document usefulness and also influences the total time spent while performing the task (Liu & Wu, 2008). This work employs user task to examine the relationship between implicit parameters and explicit ratings.

Users of an information system present their intention/interest through the formulation of input queries; however this does not adequately capture their interest (White & Kelly, 2006) and the users have to go through several iteration to get accurate search results. Therefore, there is a need to augment user query input with additional sources of information obtained explicitly or implicitly from their post-click interaction with the system in order to provide users with accurate search results based on current activity and context (White & Kelly, 2006). With respect to relevance feedback, although the explicit approach (where users of a system

* Corresponding author.
E-mail addresses: akumas@uni.coventry.ac.uk (S. Akuma), aa0535@coventry.ac.uk (R. Iqbal), c.p.jayne@rgu.ac.uk (C. Jayne), aa9536@coventry.ac.uk (F. Doctor).

http://dx.doi.org/10.1016/j.chb.2016.02.064
0747-5632/

state their opinion of the system) is commonly used for movie, music and product review, it is intrusive and alters user browsing pattern (Claypool, Le, Wased, & Brown, 2001). Explicit feedback can be replaced with non-intrusive approach that tries to infer users' interest implicitly. The inference usually takes a sequence of steps which include observing user browsing behaviour, selecting the appropriate set of implicit indicators and modelling them as source of evidence for implicit relevance feedback. These implicit indicators are obtained by observing user activity, for example through user mouse event, dwell time, keyboard event, eye tracking and psychological measures (e.g. facial expression). Indicators from dwell time, mouse and key events are commonly used at low cost whilst eye tracking indicators are expensive and can be intrusive. Eye gaze has been said to have a direct link with human cognition (Buscher, Dengel, Biedert, & Van Elst, 2012), making it the most promising implicit indicator for predicting the user's perceived relevance of web documents. It has been shown to be useful in inferring cognitive states for personalisation (Conati & Merten, 2007). The cost, configuration and portability of an eye tracker mean that it is not easily applicable to real life applications. There is therefore a need to model low cost implicit indicators as a substitution for the eye gaze indicator.

The present challenge with implicit approach is that there are no standard and acceptable methods to determine how users' activities on the web relate to their interest. This study employs the use of an instrumented web browser to capture users' implicit and explicit data from client machines and store them in a central server for further processing. In this research we investigate the consistent and predictive indicators that are frequently used by a community of users with similar interests. The indicators can then be used to build a system which learns from users' behaviour and offers recommendations to them and future users with similar information needs while minimizing the time users spend finding relevant documents.

In this paper, we investigate the relationship between implicit and explicit feedback parameters in different tasks. We derive a predictive model from implicit indicators that can be used to estimate document relevance. Finally, we validate the predictive model with eye gaze tracker. The paper is focused on answering the following research questions:

1. Can specific task situations be used to derive a predictive function model from classical implicit indicators that will signify that a web document is relevant?
2. Can eye gaze predictive indicators be substituted by a predictive function based on classical implicit indicators?

The remaining part of this paper is structured as follows: Section 2 gives an overview of related work. Section 3 presents the methodology and two user studies. Section 4 presents the results and Section 5 provides the conclusion and outlines our future work.

2. Related work

Considerable research has been carried out to improve the quality of information retrieval systems by the use of relevance feedback. Particularly the focus of research has been on implicit relevance feedback or implicit feedback. The Implicit feedback approach uses implicit indicators to replace explicit rating for the development of recommender systems (Ding, Liu, & Tao, 2010; Iqbal, Grzywaczewski, James, Doctor, & Halloran, 2012). It is used unobtrusively to infer the user's information needs based on their interest. Although implicit feedback is widely available, it is considered a secondary option to explicit feedback (Jawaheer et al., 2014) and is noisy and less accurate compared to the explicit method (Claypool et al., 2001). Current research investigates the best way of replacing explicit feedback measures with implicit feedback approaches (Alhabashneh, Iqbal, Doctor, & Amin, 2015). For instance, in a controlled setting, mouse and scroll movements have been found to exhibit some correlation with explicit rating, but it is somewhat difficult to interpret that in the real world (Buscher, Biedert, Heinesch, & Dengel, 2010). An advantage of the implicit approach is that a large amount of data can be collected ubiquitously without restricting a user to a particular place. The predictive strength of a number of implicit indicators has been investigated in the field of information retrieval. Among the implicit indicators previously investigated include: time spent on a document (also called reading time or dwell time), mouse movement, mouse distance, mouse clicks, amount of scroll movement, copy and paste, printing, highlighting, emailing and bookmarking (Akuma, 2014; Claypool et al., 2001; Iqbal, Grzywaczewski, Halloran, Doctor, & Iqbal, 2015; Konstan et al., 1997; Morita & Shinoda, 1994). Unlike explicit rating which is intrusive, expensive and alters users' browsing behaviour, implicit measures remove the cognitive cost of rating and these are not intrusive (Zemirli, 2012). The next phase of the review will focus on commonly used implicit feedback measures (classical implicit indicators) and gaze-based feedback measures.

2.1. Commonly used implicit indicators

Dwell time is one of the most researched implicit indicators. It is the duration that a document is in focus. In an attempt to effectively substitute explicit rating with implicit measures, Morita and Shinoda (Morita & Shinoda, 1994) introduced the use of dwell time as a behavioural characteristic for creating user profile and data filtration. They conducted an experiment with 8 users who were given a six week task to read articles in a news group they belonged to and explicitly rate them. The investigation was based on how the length of the document, its readability and the amount of unread article affects the reading time. They found that users spend longer time on articles they find interesting but the length of an article does not have any significant effect on the reading time. Using modified distributed software, Konstan et al (Konstan et al., 1997) repeated the study of Morita and Shinoda (Morita & Shinoda, 1994) in a natural setting. Explicit rating and reading time was logged from participants in a recommender system trial. Their findings show that a recommender system based on reading time is as accurate as an explicit recommender system. Both Morita & Shinoda, 1994 and Konstan et al (Konstan et al., 1997) research was based on a single implicit indicator (reading time) and users were restricted to certain news groups thereby limiting their 'true' web experience. In a related study with a focus on academic and professional journal articles, Kim, Oard and Romanik (Kim, Oard, & Romanik, 2000) developed a framework to investigate whether reading time is a good predictor of user interest and whether retention indicators like printing can be used to augment the predictive power of the reading time. Their findings indicate that users spent more time on documents they consider relevant. Research by Núñez-Valdéz et al (Núñez-Valdéz et al., 2012) found that display time and number of visits can be used to represent users' interest in recommending electronic books. Akuma et al (Akuma, Jayne, Iqbal, & Doctor, 2014) investigated the predictive strength of dwell time and mouse activity in a task-specific context. They correlated user generated implicit indicators with explicit relevance ratings on a set of documents and found a positive correlation between the dwell time and the explicit relevance rating. Yi et al (Yi, Hong, Zhong, Liu, & Rajan, 2014) examined the use of item-level dwell time to infer document relevance. They used both client and server-side logging to capture real-world user data from Yahoo across

different devices. They infer that dwell time can be integrated and used for personalized recommender system. Although most research suggests that dwell time is an important indicator of interest; Kelly and Belkin (Kelly & Belkin, 2004) cautioned that dwell time vary in task and subject and are difficult to interpret. This assertion was affirmed by Velayathan and Yamada (Velayathan & Yamada, 2007).

High dwell time on a document does not necessarily mean relevance because a user may leave a web document open for a long time without gazing at it. Guo and Agichtein (Guo & Agichtein, 2012) suggests that an aggregation of dwell time with other promising indicators like cursor movements and scroll can serve as a better evidence of relevance. Nichols (Nichols, 1997) introduced a list of additional observable behaviour indicators that can be a source for implicit feedback. These include: mark, reply, glimpse, query, associate, refer, repeated use, delete, save and print. Oard and Kim (Oard & Kim, 1998) grouped observable feedback behaviour into *Behaviour Category axis* and *Minimum Scope axis*. Behaviour Category axis represents the purpose of the observed behaviour and it comprises of Examine, Retain, Reference and Annotate behaviour while Minimum Scope axis encapsulates the smallest scope of the object in use. It consists of Segment, Object and Class. Kelly and Teevan (Kelly & Teevan, 2003) added "Create" to the Behavioural Category axis; it entails the behaviour associated with users when creating of new information. No quantitative measures were provided by Oard and Kim (Oard & Kim, 1998) to support the effectiveness of the indicators and only a few of the observable feedback behaviour are frequently used by online users. This work focuses on the frequently used online behaviour that can be employed to assist users to retrieve relevant web documents in an interactive information retrieval environment.

The 'Curious Browser' is a web browser developed by Claypool et al (Claypool et al., 2001) to study the predictive strength of some commonly used indicators. Implicit data were captured as users browse while the explicit data were collected through a five point rating scale. The implicit indicators measured were mouse clicks, scrolling, mouse movement and elapse time. Their results show that the time spent on a document, the amount of scroll and the combination of time spent and amount of scroll are good predictive indicators which have a stronger correlation with user explicit ratings. Kim and Chan (Kim & Chan, 2005) examined similar indicators to Claypool et al. (Claypool et al., 2001). In their study, participants were asked to save over 5 pages, bookmark more than 10 pages, print over 5 pages and use 'memo' on more than 5 pages. They found that distance of mouse movement and time spent are the promising indicators of interest. Both Claypool et al (Claypool et al., 2001) and Kim et al (Kim & Chan, 2005) did not use information seeking tasks to engage participants; this might create uncertainty in applying the findings. Also, Kim & Chan compelled the users to behave in some way (bookmark, print, save) which might have affected their 'true' web experience.

Zhu, He and Wang (Zhu, He, & Wang, 2012) used the parameters of Clicks, Bookmaking, Voting and Reply to adaptively model users' interest. Fox et al. (Fox, Karnawat, Mydland, Dumais, & White, 2005) correlated implicit and explicit judgement and developed a predictive model using the Bayesian method. They found that the aggregation of time spent on the search result page, click-through, and how users ended a session or exited a result page, gave the best prediction for their explicit judgement of satisfaction. The study by Fox et al (Fox et al., 2005) focused on search engine result page and not user's post-click behaviour on documents, which is the focus of this work. In the context of electronic book recommendation, Núñez-Valdez et al (Núñez-Valdez, Lovelle, Hernández, Fuente, & Labra-Gayo, 2015) proposed an architecture that analysed and transformed implicit feedback parameters to approximate explicit ratings for a community of readers. Their results show that users' interest can be determined by analysing and converting their behaviour. Leiva and Huang (Leiva & Huang, 2015) used a client-side approach of tracking user activity to record users' cursor movements for computing relevance of search results. They infer that the 'cursor movement' capturing tool is a viable tool for understanding user behaviour.

Zemirli (Zemirli, 2012) worked on post-retrieval documents and he developed a web browser (WebCap) that uses 'examine and retention' indicators to infer users' interest in real time. He conducted an experiment on 6 users and found that WebCap was able to capture 80% of relevant documents when compared with explicit user judgements. Similar success was reported by Shapira, Taieb-Maimon and Moskowitz (Shapira, Taieb-Maimon, & Moskowitz, 2006). Balakrishnan and Zhang (Balakrishnan & Zhang, 2014) examined the effect of some implicit indicators on post-retrieval document relevancy. They found that a combination of text selection, dwell time, click-through and page review post-click behaviour can improve the precision of relevance feedback.

Most studies of implicit indicators in the context of information retrieval have been based on Search Engine Result Page (SERP) and very few studies have focussed on developing a predictive model through the aggregation of implicit indicators generated from web documents visited by users. This work uses a task based approach and it focuses on user post-click behaviour. It revisits some of the implicit indicators used in the studies above but instead of the multiple domain approach of data collection, information tasks are used to limit users' goals in a particular domain. The following implicit indicators are evaluated in this work: users' dwell time, mouse movement, mouse distance, mouse velocity, mouse clicks, amount of scroll, keystroke and amount of copy. A stepwise regression is utilised to extract and model the most predictive indicators. Investigation is also carried out to examine whether the model can be used in place of eye gaze measures.

2.2. Eye gaze-based implicit indicators

Research in eye gaze for information retrieval tasks has shown that there is a direct link with human cognition, making it the most promising indicator for predicting users' perceived relevance on web documents. The modern eye trackers have a better degree of accuracy and precision in measuring gaze features as compared to previous ones. This has led to an increased study of gaze parameters as they relate to information retrieval (Gwizdka, 2014). A previous study by Buscher et al (Buscher et al., 2012) suggests that eye gaze is an important indicator of interest and it has a direct link to a user's visual attention. In a related study conducted by Salojärvi et al (Salojärvi, Puolamäki, & Kaski, 2005) to explore whether eye movement can be used as a source of implicit relevance information, they found that accurate prediction of document relevance can be deduced from eye movement. Cole et al (Cole et al., 2011) investigated how user eye tracking information can be used to obtain user intent and interest during information retrieval. They found that eye movement during reading is a good implicit indicator for a user's task type. Granka, Joachims and Gay (Granka, Joachims, & Gay, 2004) also reported that eye movement a good indicator of interest.

Although gaze features are said to be the most predictive indicator of interest (Buscher et al., 2012), they are however not used in the real world due to the expensive cost of an Eye tracker, its configuration and portability. It is therefore necessary to substitute the gaze features that has direct link to human cognition with other implicit indicators obtained from 'cheap and available' sources. Attempts have been made by researchers to substitute the eye gaze indictor with other single implicit indicators. Huang, White and

Dumais (Huang, White, & Dumais, 2011) found a slight coordination between cursor movement and eye gaze. Guo and Agichtein (Guo & Agichtein, 2010) say that regions where mouse pointer and eye are within 100 pixels of each other can be predicted accurately by nearly 77%. Most of the previous studies focused on finding a relationship between mouse cursor and eye gaze on Search Engine Result Page (SERP). In this work, we examined how we can validate the strength of the predictive function derived from aggregating commonly used implicit indicators with eye gaze. We show that to a reasonable degree, we can use an aggregation of non-gaze implicit indicators as a substitution to gaze-based indicators.

3. Methodology and user studies

Two user studies were conducted. The first user study "Classical Implicit Indicators vs explicit ratings" analyses 'commonly available' implicit feedback measures in relation to explicit relevance ratings for specific information retrieval tasks. It correlates implicit and explicit feedback parameters, and it uses multilinear regression to derive a predictive model that can estimate document relevance. The second user study "Eye Gaze Indicators vs explicit ratings" validates the derived predictive model with the eye gaze measure.

3.1. User study 1 method (classical implicit indicators vs explicit ratings)

The goal of this study is to investigate the relationship between classical implicit indicators and explicit relevance ratings in task specific domain, and to identify the implicit indicators which have a stronger relationship with the explicit ratings to use in the predictive model. Since there is no agreeable standard for sample size for an Interactive Information Retrieval (IIR) experiment and no linear relationship between sample and population (Kelly, 2009), we worked with a sample size that is consistent with previous research. The participants recruited for study 1 were 77 undergraduate students of Computing. Only participants above 18 years were allowed to participate. The participants recruited for this study had a high proficiency with the use of computers. Remuneration was not given to the participants but they were informed of the overall benefit of the research to enhance student learning through the recommendation of documents by using relevance feedback approaches.

The participants were asked to perform search tasks and produce a short report of their findings. The study took 45 min for each of the group and the participants were given an option to either perform the experiment in a controlled environment or at home in a natural setting. The controlled study was performed in selected computer laboratories of the university. A consent form was given to them to complete which was followed by a short tutorial about the task. A zipped Mozilla Firefox Portable with an embedded JavaScript plugin was uploaded to Google drive for participants to download and extract to their computers. After the extraction, they were instructed to open the Firefox browser in the folder and begin the search for answers to the given task. The participants were informed not to look at their clock while performing the task; this was to remove any anxiety and nervousness. Participants either entered their own query in a search engine to find their required documents or they entered the URL address of their required web documents directly. For every web document they visited, participants were prompted to enter their User Id and then do the following:

- Read the document for information relating to their task then close the current tab (web document). On closing the tab, participants were asked to explicitly rate the documents according to its relevance.
- The participants were required to visit and read a minimum of 7 web documents
- Finally, participants were asked to prepare a 200 words report of the solution to the given task.

The injected JavaScript plugin captured users' behavioural features (see Table 1) and explicit ratings for relevance for each document visited. URLs from Google results pages, Facebook and Yahoo were prevented from being logged on to ensure that only documents related to the task at hand were recorded. The data collected were then sent to a central server, and then to a MySQL database for storage. Fig. 1 depicts how data were automatically captured and stored while the users were performing their tasks.

3.1.1. Description of tasks

In this section, we discuss the tasks which were given to the students. The tasks employed for this research are simulated work situations which follow the recommendation by Borlund (Borlund, 2003). The tasks are designed to encourage the participants to search the web naturally. They are designed with the intention to be interesting and relatable for the participants. The tasks domain for the study is in the area of Computer Science. Two simulated search tasks are composed of two parts – the *simulated work task situation* and *indicative request*. The simulated work task situation is a 'cover story' that has three characteristics:

- The subjects should be able to relate and identify with the task
- The subjects should find the topic situation interesting
- The task situation should have an imaginative context so that subjects can relate to it.

When tasks are assigned to users without a background information or context, it demotivates the users and they may consider the task as artificial (Kelly, 2009). A classification scheme by Liu, Liu and Belkin (Liu, Liu, & Belkin, 2013) is used to group the task components into a single scheme as shown in Table 2.

3.1.1.1. Mixed task (task 1). Task 1 is considered a mixed product (Decision and Intellectual task) because it involves making a decision to solve a problem with the most efficient method (RUP or Waterfall Model). It also focuses on 'how' a problem can be solved. It asks for the most important stage of the lifecycle, making it also an intellectual task. The goal of the task is specific because participants have to find a particular approach and it is also of high complexity because at least 7 documents have to be sourced.

3.1.1.2. Simulated work task situation 1. GIG Software Development company employed you as a consultant to provide a solution to the Company's pressing problem of developing a customized software within a minimal time frame. Some professional software developers achieved this by using the Rational Unified Process while others used the waterfall model.

3.1.1.3. Indicative request 1. Which approach would you consider for a small project of few lines of code (LOC) and what stage of the software life cycle would you consider to be the most important? State the reasons for your answer in your report.

3.1.1.4. Factual task (task 2). Task 2 is considered factual because specific facts have to be sourced. It focuses on gathering information on a subject and participants have to find specific information which is explicitly measurable. The complexity for this task is high because at least 7 web documents have to be retrieved.

Table 1
Implicit and explicit parameters captured.

Parameters	Description
Dwell Time (DT):	This is the accumulated time in seconds spent by a user on an active page during browsing. It is also called reading time.
Distance of Mouse Movement (DMM)	The Euclidean distance of mouse movement is calculated by its X and Y coordinates on the monitor in every 100 ms.
Total Mouse Movement (TMM):	This is the total mouse movement calculated by its X and Y coordinates on the monitor. The count is incremented by one for X and Y as the mouse hovers.
Mean Mouse Velocity (MMV)	This is the total speed covered by the mouse on the monitor.
Number of Mouse Clicks (NMC)	This is the total amount of mouse clicks on a page. The number of mouse click is incremented every time the mouse is clicked by a user.
Amount of Scroll (AS)	Most web pages are longer in length than the monitor height. When readers are interested in a page, they scroll the page. The scrolling is normally done by either clicking or dragging the scroll bar. Any time a user clicks the scrollbar up or down, the count is incremented.
Number of Keystrokes (NK):	This is the total number of keystrokes on a document. This is incremented when the user strikes a key.
Amount of Copy (AC)	This is the number of times text is copied to the clipboard from a document. It is incremented by one any time text from a particular document is copied.
Mouse Duration Count (DC)	This is the total number of 100 ms intervals that occurs while the mouse is moved on the screen.
Time Stamp	This is the time and date in GMT when a document is loaded and when a document is closed.
URL	This is the http address of any web document visited by a user.
IP Address (IP)	This is the internet protocol address of a user. It represents the user's location.
Explicit Relevance Ratings (ER)	This is the actual rating of the web document by the user. The Firefox plugin attaches a six scale rating button on each of the webpages. After reading a webpage, the user rates it by clicking on any of six scale buttons where 5 – means very relevant, 4 – means more relevant, 3 – means moderate relevant, 2 – means slightly relevant, 1– means very low relevance, 0 – means not relevant.

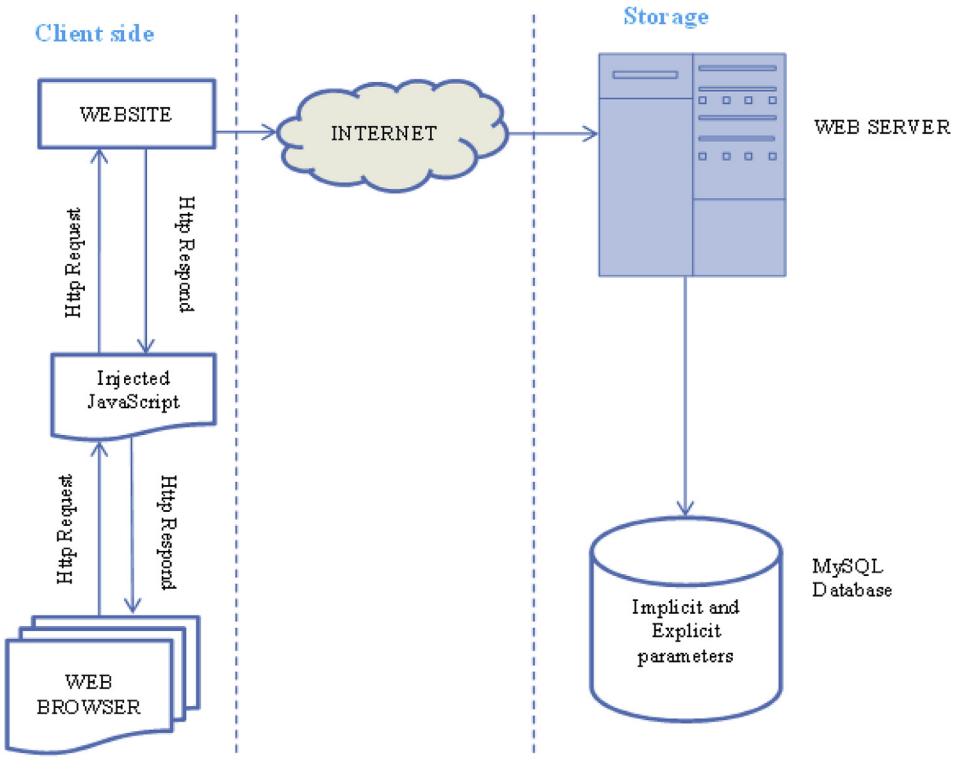


Fig. 1. The experimental system used for capturing data.

3.1.1.5. Simulated work task situation 2. Google is looking for young and ambitious students of Computer science for internship to work under the Company's Service Management Department. Consider that you are shortlisted for an interview among 2000 applicants and you are asked to search the internet and find answers to questions related to Information Technology Infrastructure Library (ITIL):

Table 2
Component grouping of the two tasks.

Task	Product	Goal (quality)	Objective complexity
Task 1	Mixed	Specific goal	High
Task 2	Factual	Specific goal	High

3.1.1.6. Indicative request 2

a) What are the five stages of the ITIL lifecycle?

b) What are the differences between ITIL v1, v2 and v3 (2007)?

c) What are ITIL processes?

d) What are ITIL functions?

e) Who should use ITIL?

f) When should ITIL be used?

g) What are the differences between ITIL and ISO/IEC?

3.2. User study 2 method (Eye Gaze Indicators vs explicit ratings)

The purpose of this study is to correlate eye gaze measures with explicit relevance ratings on some web documents to validate the predictive model which is developed as result of user study 1. Briefly, based on the predictive model, the documents perceived to be the most relevant and least relevant are identified from the pool of documents. These documents are then given to participants to read through with an eye tracker installed on the machine and rate them according to how relevant they are to the task under consideration. The rating categorisation is the same as that employed in user study 1. In this user study, 9 university students majoring in the area of Computer Science took part. A task brief explaining the procedure and a consent form was given to the participants to complete. This was followed by a short tutorial about the experiment. The eye tracker was calibrated on a five point calibration scale and each of the participants had to sequentially read through the 6 documents within 30 min and rate them.

3.2.1. Apparatus

Gaze data were captured with a Tobii TX300 desk mounted eye-tracker. It was paired with a 23 inch LCD monitor with a resolution of 1920 × 1080 pixels. The tracking frequency for the eye tracker was 300 Hz and it gave room for subjects to move their heads. The accuracy was 0.4 degree of visual angle. The participants' fixation count, fixation duration and heat map as shown in Table 3, were captured by Tobii SDK Software.

4. Results

This section presents the results of user study 1 and user study 2. Section 4.1 presents the results of study 1 and the results of study 2 presented in section 4.2.

4.1. Relationship between implicit indicators and explicit relevance ratings

Every participant in user study 1 visited at least one web documents while performing the assigned tasks. The maximum number of websites visited by a participant in a single task was 11 and the minimum was one. A total number of 343 web pages were analysed. Pearson correlation is employed in finding the correlation between the user explicit ratings and the implicit indicators. In order to determine if results from Pearson correlation are real and not random, significance testing is employed. A confidence interval of 95% and a statistical significant coefficient, $p < 0.05$, is used for analysing the dataset as explained in Table 4.

Null hypothesis indicates that there is no statistical significant relationship or association between two measured parameters. When the null hypothesis is rejected, it means there is evidence that there is a relationship between two parameters.

Initial analysis for linearity shows that the implicit variables were linearly related with the explicit relevance ratings. Table 5 illustrates the Pearson correlation between the implicit indicators and explicit relevance ratings. Positive correlation was observed between explicit relevance rating and the following implicit indicators: the number of mouse clicks, amount of copy, the amount of scroll, the mouse movement along X-axis and Y-axis, the dwell time, the mouse distance and the mouse duration count. There was no significant correlation between the explicit relevance ratings and the mean mouse speed, and between the explicit relevance ratings and the keystroke. The correlation coefficient between implicit indicators and explicit relevance ratings obtained in this work is higher than those obtained in previous research (e.g. Guo and Agichtein (Guo & Agichtein, 2012)) conducted under similar context.

4.1.1. Consistency in explicit relevance rating

Since the goal of this research is to obtain common consistent implicit indicators that can be used to represent users' interest, the consistency of the participants' explicit relevance ratings was examined. Common documents visited by the users were extracted from the pool of documents visited. Among these common documents extracted, the most viewed document had 21 hits while the least had 2 hits. An investigation was carried out to see if documents commonly visited are relevant. The mean of the explicit relevance ratings of the common documents visited was found to be 3.21. Since '3' was labelled as 'relevant' on our 6 point subjective explicit rating scale, it therefore indicates that common documents visited by users are relevant across the users. It is therefore concluded that common consistent implicit indicators captured across all user populations can be used to infer relevance among a community of users. The findings of user study 1 led to the development of a predictive model.

4.1.2. Predictive model

To aggregate the most predictive indicators of interest for a function model, multiple linear regression analysis was used to estimate the user behavioural features that best represent the user explicit interest. A stepwise regression method was employed for this selection. The stepwise regression automatically selects the predictive variables through a sequence of t-test. The regression function is given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_n X_n \quad (1)$$

where

$\beta_0, \beta_1 \ldots \beta_n$ are regression coefficients (unknown model parameters). Y is the dependent variable while X_1, X_2 and X_n are independent variables.

Table 3
Gaze measures captured.

Parameters	Description
Total Fixation Duration (TFD):	This is the sum of duration of all individual fixations within a specific area of interest of a document. Individual fixation is between 250 ms and 300 ms.
Total Fixation Count (TFC)	This is number of times that a user fixates within a specific area of interests of a document.
Heat Map	This is a visualization technique that separates different levels of fixation intensity, it show areas that are more fixated to be denser than areas that are less fixated.

Table 4
Statistic testing and condition for the result of Pearson correlation to be significant.

Statistic significant test (p)	Coefficient level	Condition	Action
Pearson correlation	$> = 0.05$	There is no significant relationship	Accept null hypothesis
	<0.05	There is a significant relationship	Reject null hypothesis

Table 5
Pearson correlation between the implicit indicators and explicit relevance ratings.

Implicit Indicators	Pearson correlation (r) with user explicit rating	Significant coefficient level (p)
Number of Clicks	0.211	0.000
Amount of Copy	0.286	0.000
Amount of Scroll	0.123	0.023
Mouse movement X	0.225	0.000
Mouse movement Y	0.261	0.000
Dwell time	0.285	0.000
Mouse distance	0.254	0.000
Mouse duration count	0.238	0.000
Mean mouse speed	−0.73	0.180
Keystroke	−0.18	0.742

Among the nine features entered as inputs, the stepwise regression included only dwell time and amount of copy parameters. The predictive function below was obtained:

Explicit Relevance ratings = 2.978 + 0.281(Total Copy) + 0.002(Dwell Time).

Correlation coefficient 0.36.

The relationship between the explicit ratings and the implicit indicators (dwell time and amount of copy) in the predictive model have a higher correlation as compared to other features examined and only these features are sufficient to derive the predictive model. Considering that the relationship between the explicit and implicit indicators in the predictive model produced a correlation coefficient of 0.36, which is higher than that of the individual indicators examined, it suggests that when the features are aggregated, a higher degree of accuracy in prediction is obtained.

To conduct a study to validate the predictive model, an additional column was created for the predictive model in the dataset containing web documents and their related implicit generated indicators. The predictive function was used to compute a score for each of the documents. The computed score was sorted in a descending order and documents perceived to be highly relevant and least relevant were extracted from the pool of documents. This set of documents was then presented in a sequential order to participants in user study 2 to read before an eye tracker and rate their relevance.

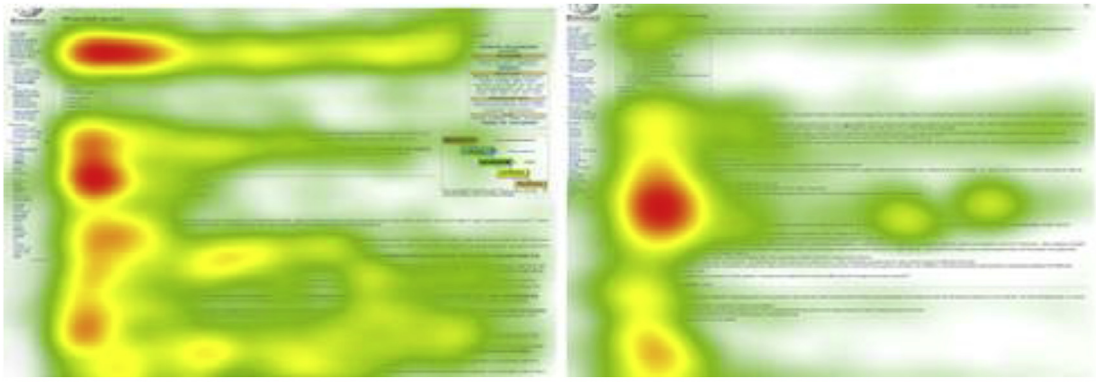
4.2. Validation of the predictive model based on "Eye Gaze Indicators" user study

The results of user study 1 suggest that user interest on web documents can be inferred from their behavioural activity. The aim of user study 2 is to use eye gaze measures (Fixation count, Fixation duration, Heat map) to validate the predictive strength of the function model derived in 4.1.2, and to show that to a certain degree of accuracy, the model can be used in place of an eye gaze. The data captured by the eye tracker from one of the participants (participant 6 in Table 6) was excluded from the analysis due to poor calibration which led to incomplete data. Only the remaining 8 participants' results were analysed. The correlation between total fixation duration and explicit ratings was not statistically significant. This is consistent with the findings of Buscher et al (Buscher et al., 2012). There was however a statistical significant correlation of 0.32 ($p = 0.025$) between the total fixation count and user explicit ratings. Qualitative results from the heat maps (Fig. 2) show that documents explicitly rated highly relevant were denser and had higher mean fixation count than documents rated as low relevance. Fig. 2 shows the heat maps for the document with the highest mean fixation count and the document with lowest mean fixation count.

The results in user study 2 show a significant correlation between the fixation count and the user perception of relevance. There was no significant correlation between fixation duration and user ratings. However, the explicit ratings for the documents identified in user study 1 were correlated with the mean explicit ratings of the documents in user study 2 and a strong correlation of

Table 6
The recordings for the Mean Fixation Count (MFC) and Explicit Ratings (ER).

Participants	URL1		URL 2		URL 3		URL 4		URL 5		URL 6	
Parameters	TFD1	ER1	TFD2	ER2	TFD3	ER3	TFD4	ER4	TFD5	ER5	TFD6	ER6
Rec 01	117.0	4	583.0	5	166.0	4	127.0	4	37.0	3	297.0	4
Rec 02	361.0	4	359.0	4	137.0	5	68.0	2	21.0	3	298.0	2
Rec 03	542.0	4	531.0	4	226.0	3	239.0	4	72.0	3	348.0	2
Rec 04	12.0	4	163.0	5	172.0	5	26.0	2	8.0	5	31.0	2
Rec 05	5.0	2	614.0	4	121.0	2	9.0	2	13.0	2	180.0	1
Rec 06	68.0	4	12.0	4	0.0	3	1.0	3	0.0	2	0.0	3
Rec 07	2325.0	5	360.0	4	171.0	4	554.0	4	163.0	5	641.0	3
Rec 08	1523.0	4	2115.0	4	836.0	3	485.0	5	809.0	4	1407.0	4
Rec 09	622.0	3	1484.0	4	236.0	3	61.0	2	189.0	4	403.0	3
Mean Recordings	619.4	3.75	691.2	4.25	229.4	3.625	174.4	3.125	145.8	3.625	400.6	2.625



a) Highest mean fixation count b) Lowest mean fixation count

Fig. 2. The Heat maps of documents with the highest mean fixation count and the lowest mean fixation count.

Table 7
Comparison of study 1 and study 2 result.

Parameters	Predictive Model	Explicit ratings (User study 2)
Fixation count		0.32
Explicit ratings (user study 1)	0.36	0.82

0.82 ($p = 0.045$) was obtained, showing consistency in the ratings of the documents by the participants in the two user studies. The results of the correlation obtained in the two user studies suggest the predictive model can be used in place of Fixation count when an eye tracker is not available. The correlation between the explicit ratings of the predictive function model used for identifying and extracting the documents from the dataset is 0.36. The correlation between the total fixation count and the explicit rating is 0.32. Considering the consistency in the ratings of the documents in the two studies, and the predictive model and fixation count producing similar correlation coefficient with the explicit ratings (as shown in Table 7), we can infer that there is no significant difference between the predictive model based on implicit indicators and the eye gaze in the context employed. The relationship between the ratings is shown in Fig. 3. The predictive model derived can be used to estimate document relevance for the recommendation of relevant web documents to a set of users based on their previous interaction with the system.

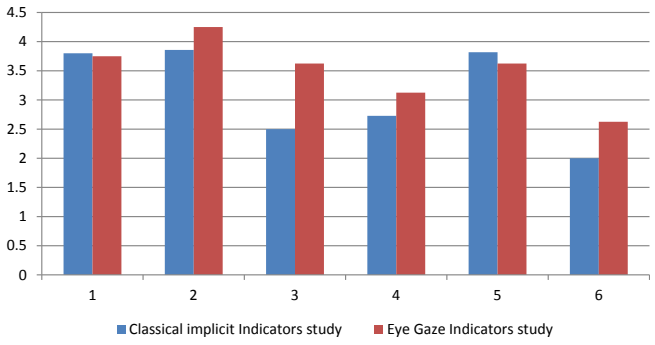


Fig. 3. Graph showing the explicit ratings of users in study 1 and the mean explicit ratings of users in study 2.

5. Conclusions

This work employed a task context approach to explore the relationship between implicit and explicit user parameters. Our findings show that apart from keypress and mean mouse speed, other classical implicit indicators measured correlate with the explicit ratings. A predictive function model was derived from the relationship between implicit indicators and explicit ratings. The model consisted of two of the most predictive implicit indicators (Dwell time and Copy) and it had a higher correlation coefficient than the single indicators. A validation study based on eye gaze was used to confirm the predictive strength of the predictive model and it showed that there was no significant difference between the predictive model based on implicit indicators and the eye gaze in predicting perceived relevant documents within the context examined. These findings provide a cost effective method for understanding user behaviour in a task specific context through the use of implicit indicators which can be used in education-based recommender systems.

The data collection in user study 1 was based on capturing documents that were JavaScript enabled. Pdf documents, images and video web resources were not captured. This is a limitation that will be addressed in our future work. Another limitation in this work is that only selected documents were identified and used to examine the relationship between the predictive function model and gaze based measures. A better approach would be to capture the eye gaze measures and the classical implicit indicators simultaneously in a single study. The eye tracker used in this work was limited in the searching procedure. The Tobii SDK Software used for capturing gaze measures is only compactible with Internet explorer browser while the JavaScript plugin designed to capture the classical indicators is specific to Mozilla Firefox browser. A cross-browser plugin to capture classical indicators will be developed for future work.

Future work will include the development of a framework for a context-based recommender system to improve post-retrieval document relevancy. The framework will use Vector Space Model (VSM) to index and rank documents based on query-document similarity, and the predictive model derived from users' previous interaction with the system will be used to re-rank the documents according to users' interest/perceived relevance.

References

Akuma, S. (2014). Investigating the effect of implicit browsing behaviour on students' performance in a task specific context. *International Journal of*

Information Technology and Computer Science(IJITCS), 6, 11–17.
Akuma, S., Jayne, C., Iqbal, R., & Doctor, F. (2014). Implicit predictive indicators: mouse activity and dwell time. In *10th IFIP WG 12.5 international conference, AIAI 2014, Rhodes, Greece* (pp. 162–171).
Alhabashneh, O., Iqbal, R., Doctor, F., & Amin, S. (2015). Adaptive information retrieval system based on fuzzy profiling. In *IEEE International Conference on Fuzzy Systems, Istanbul, Turkey*.
Alhindi, A., Kruschwitz, U., Fox, C., & Albakour, M. (2015). Profile-based summarisation for web site navigation. *ACM Transactions on Information Systems, 33*, 1–40.
Balakrishnan, V., & Zhang, X. (2014). Implicit user behaviours to improve post-retrieval document relevancy. *Computers in Human Behavior, 33*, 104–112.
Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research, 8*.
Buscher, G., Biedert, R., Heinesch, D., & Dengel, A. (2010). Eye tracking analysis of preferred reading regions on the screen. In *Conference on human factors in computing systems – Proceedings* (pp. 3307–3312).
Buscher, G., Dengel, A., Biedert, R., & Van Elst, L. (2012). Attentive documents: eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems, 2*, 1–30.
Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In *International conference on intelligent user interfaces, proceedings IUI* (pp. 33–40).
Cole, M. J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N. J., & Zhang, X. (2011). Task and user effects on reading patterns in information search. *Interacting with Computers, 23*, 346–362.
Conati, C., & Merten, C. (2007). Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. *Knowledge Based System, 20*(8), 557–574.
Ding, L., Liu, B., & Tao, Q. (2010). Hybrid filtering recommendation in e-learning environment. In *2nd international workshop on education technology and computer science, ETCS 2010* (pp. 177–180).
Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve Web search. *ACM Transactions on Information Systems, 23*, 147–168.
Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *Proceedings of Sheffield SIGIR – Twenty-seventh annual international ACM SIGIR conference on research and development in information retrieval* (pp. 478–479).
Grzywaczewski, A., & Iqbal, R. (2012). Task-specific information retrieval systems for software engineers. *Journal of Computer and System Sciences, 78*(4), 1204–1218. Elsevier.
Guo, Q., & Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. In *Conference on human factors in computing systems – Proceedings* (pp. 3601–3606).
Guo, Q., & Agichtein, E. (2012). Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In *WWW'12-Proceedings of the 21st annual conference on world wide web* (pp. 569–578).
Gwizdka, J. (2014). Characterizing relevance with eye-tracking measures. In *Proceedings of the 5th information interaction in context symposium, IIiX 2014* (pp. 58–67).
Huang, J., White, R. W., & Dumais, S. (2011). No clicks, no problem: using cursor movements to understand and improve search. In *Conference on human factors in computing systems - Proceedings* (pp. 1225–1234).
Iqbal, R., Grzywaczewski, A., Halloran, J., Doctor, F., & Iqbal, K. (2015). *Design implications for task-specific search utilities for retrieval and reengineering of code,Enterprise Information Systems*, 1751–7575. Taylor and Francis. http://dx.doi.org/10.1080/17517575.2015.1086494.
Iqbal, R., Grzywaczewski, A., James, A., Doctor, F., & Halloran, J. (2012). Investigating the value of retention actions as a source of relevance information in the software development environment. In *Proceedings of the 2012 IEEE 16th international conference on computer supported cooperative work in design, CSCWD 2012* (pp. 121–127).
Jawaheer, G., Weller, P., & Kostkova, P. (2014). Modeling user preferences in recommender systems: a classification framework for explicit and implicit user feedback,". *ACM Transactions on Interactive Intelligent Systems, 4*, 1–26.
Kellar, M., Watters, C., Duffy, J., & Shepherd, M. (2004). Effect of task on time spent reading as an implicit measure of interest. *Proceedings of the ASIST Annual Meeting, 41*, 168–175.
Kellar, M., Watters, C., & Shepherd, M. (2007). A field study characterizing Web-based information-seeking tasks. *Journal of American. Society for Information Science and Technology, 58*, 999–1018.
Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval, 3*, 1–224.
Kelly, D., & Belkin, N. J. (2004). Display time as implicit feedback: understanding task effects. In *Proceedings of Sheffield SIGIR - Twenty-seventh annual international ACM SIGIR conference on research and development in information retrieval* (pp. 377–384).
Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference. *SIGIR Forum, 37*, 18–28.
Kim, H. R., & Chan, P. K. (2005). Implicit indicators for interesting web pages. In *WEBIST 2005-1st international conference on web information systems and technologies, proceedings* (pp. 270–277).
Kim, J., Oard, J. D. W., & Romanik, K. (2000). *User modeling for information access based on implicit feedback*. University of Maryland, College Park. Technical Report: HCIL-TR-2000-11/UMIACS-TR-2000-29/CS-TR-4136.
Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Applying collaborative filtering to usenet news. *Communications ACM, 40*, 77–87.
Leiva, L. A., & Huang, J. (2015). Building a better mousetrap: compressing mouse cursor activity for web analytics. *Information Processing & Management, 51*(3), 114–129.
Li, Y., & Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management, 44*, 1822–1837.
Liu, J., Liu, C., & Belkin, N. (2013). Examining the effects of task topic familiarity on searchers' behaviors in different task types. *Proceedings of the ASIST Annual Meeting, 50*.
Liu, D., & Wu, I. (2008). Collaborative relevance assessment for task-based knowledge support. *Decision Support System, 44*, 524–543.
Morita, M., & Shinoda, Y. (1994). Information filtering based on user behaviour analysis and best MatchText retrieval. In *In proceedings of SIGIR conference on research and development* (pp. 272–281).
Nichols, D. M. (1997). Implicit ratings and riltering. In *In proceedings of the 5th DELOS workshop on filtering and collaborative filtering, budapaest, Hungary, ERCIM* (pp. 10–12).
Núñez-Valdéz, E. R., Cueva Lovelle, J. M., Sanjuán Martínez, O., García-Díaz, V., Ordoñez De Pablos, P., & Montenegro Marín, C. E. (2012). Implicit feedback techniques on recommender systems applied to electronic books. *Computers in Human Behavior, 28*, 1186–1193.
Núñez-Valdez, E. R., Lovelle, J. M. C., Hernández, G. I., Fuente, A. J., & Labra-Gayo, J. E. (2015). Creating recommendations on electronic books: a collaborative learning implicit approach. *Computers in Human Behavior, 51*, 1320–1330.
Oard, D., & Kim, J. (1998). Implicit feedback for recommendation systems. In *In proceedings of the AAAI workshop on recommender systems* (pp. 81–83).
Salojärvi, J., Puolamäki, K., & Kaski, S. (2005). Implicit relevance feedback from eye movements. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3696*, 513–518. LNCS.
Shapira, B., Taieb-Maimon, M., & Moskowitz, A. (2006). Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. In *Proceedings of the ACM symposium on applied computing* (pp. 1118–1119).
Velayathan, G., & Yamada, S. (2007). Behavior-based web page evaluation. In *Proceedings – 2006 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT 2006 workshops proceedings)* (pp. 409–412).
White, R. W., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *International conference on information and knowledge management, proceedings* (pp. 297–306).
Yi, X., Hong, L., Zhong, E., Liu, N. N., & Rajan, S. (2014). Beyond clicks: dwell time for personalization. In *RecSys 2014-Proceedings of the 8th ACM conference on recommender systems* (pp. 113–120).
Zemirli, N. (2012). WebCap: Inferring the user's interests based on a real-time implicit feedback. In *7th international conference on digital information management, ICDIM 2012* (pp. 62–67).
Zhu, Y., He, L., & Wang, X. (2012). User interest modeling and self-adaptive update using relevance feedback technology. *Procedia Engineering*, 721–725.