User Model User-Adap Inter (2009) 19:307–339
DOI 10.1007/s11257-009-9066-4

ORIGINAL PAPER

Can eyes reveal interest? Implicit queries from gaze patterns

Antti Ajanki · David R. Hardoon · Samuel Kaski · Kai Puolamäki · John Shawe-Taylor

Received: 7 July 2007 / Accepted: 20 August 2009 / Published online: 10 September 2009

Abstract We study a new research problem, where an implicit information retrieval query is inferred from eye movements measured when the user is reading, and used to retrieve new documents. In the training phase, the user's interest is known, and we learn a mapping from how the user looks at a term to the role of the term in the implicit query. Assuming the mapping is universal, that is, the same for all queries in a given domain, we can use it to construct queries even for new topics for which no learning data is available. We constructed a controlled experimental setting to show that when the system has no prior information as to what the user is searching, the eye movements help significantly in the search. This is the case in a proactive search, for instance, where the system monitors the reading behaviour of the user in a new topic. In contrast, during a search or reading session where the set of inspected documents is biased towards being relevant, a stronger strategy is to search for content-wise similar documents than to use the eye movements.

Keywords Eye movements · Implicit relevance feedback · Information retrieval · Machine learning · Support vector machines

The authors appear in alphabetical order.

A. Ajanki (✉) · S. Kaski · K. Puolamäki
Department of Information and Computer Science, Helsinki Institute for Information Technology, Helsinki University of Technology (TKK), P.O. Box 5400, 02015 Espoo, Finland
e-mail: antti.ajanki@tkk.fi

D. R. Hardoon · J. Shawe-Taylor
Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK

1 Introduction

Current information retrieval (IR) systems rely mostly on explicit, typed queries, combined with explicit feedback telling the system which of the search results were relevant. The relevance feedback is used to refine the query, and the search converges iteratively towards more relevant documents. The standard web search engines are simplified versions of this scheme; they take advantage of the large scale which allows inferring general relevance of documents from link data.

A main problem of this traditional IR paradigm is that formulating good textual queries is a challenging task even for experienced users (Turpin and Scholer 2006). Moreover, query-based searches are only possible if the user knows her information need. The need may also be *tacit*; there may exist very useful documents that the user does not even try to search, or in a milder form the true interest or information need may be ambiguous to the users. In all these cases it is difficult or impossible to formulate a query explicitly. It would be ideal if the system could infer the interests of the users while they work, and then have some suggestions readily available when the users ask for help. We call this task *proactive information retrieval*.

A proactive information retrieval system would additionally solve the problem that giving explicit feedback is laborious. Such a system would use *implicit feedback* to infer relevance. Several forms of implicit feedback have been used (Kelly and Teevan 2003), of which at least click-stream data, time spent during reading, amount of scrolling, and exit behaviour have been found to help in predicting explicit feedback ratings (Claypool et al. 2001; Fox et al. 2005; Joachims et al. 2005). While these sources are readily available and useful, they offer limited information about a user's interests and intentions, and the predictions are far from perfect. Hence it is important to continue searching for new sources of feedback that could be used to complement the existing ones.

Our suggestion is to use implicit feedback from the observed gaze pattern as an alternative or complementary source to infer the users' intentions. Eye tracking has already been shown to be useful in user modelling: in inferring cognitive states, traits, and performance of the user for personalization purposes and for interaction adaptation (Conati and Mertena 2007, and review of earlier works therein). Moreover, eye movements have been shown to be useful in inferring relevance of documents to be used as relevance feedback (Puolamäki et al. 2005). Hence it is imaginable that eye tracking would be useful in inferring other, even more subtle cues about user interests.

We combine the eye movements with the textual content of the documents in a novel way: *we use the eye movements to formulate an IR query*, which is then used to rank unseen documents with respect to their relevance to the current interests of the user.

In this paper we study the feasibility of this approach. The practical motivation is that eye tracking equipment is becoming cheaper and smaller, and eye tracking data is soon expected to be cheaply available for most applications. If the data turns out to be useful it is then sensible to use eye tracking recordings to complement other sources. The more exciting motivation is that eye tracking data may provide more subtle cues about the users' interests compared to lower level, time-based features, as has been found in studies of users' meta-cognition (Conati and Mertena 2007).

Since we expect the eye movement patterns to be very noisy, we study two approaches. The more challenging task is to (1) construct a query from eye movements alone. The easier one is to (2) construct a query by combining information from implicit relevance feedback from eye movements *and* explicit relevance feedback.

Implicit queries have earlier been constructed based solely on the texts the user is working on (Czerwinski et al. 1999; Dumais et al. 2004); in this work we combine this research tradition with more focused eye tracking-based inferences on which *parts* of the document are interesting. This is a feasibility study investigating whether eye tracking gives valuable information in this task. If it does, the methods can then be optimized and tailored in later studies.

The texts the user is working on could alternatively be interpreted as the *context* of the user and used to complement explicit queries (Budzik and Hammond 2000). In this work we assume explicit queries are not available, as in earlier works on implicit queries. This setting is commonplace when browsing interesting documents without a clear goal, and in the beginning of a more focused browsing session. If explicit queries are available, the implicit queries inferred from eye tracking could naturally be used to complement them according to existing principles.

To test the new approach, we devise a controlled experimental setting, in which test subjects read through short text snippets searching for documents related to a given topic. During the reading the users' eye movements are recorded with an eye tracking device. We extract term-specific eye movement features for each document the user reads. The features are then used for predicting importance of the term for the search task. The setup is an extended version of our earlier work (Puolamäki et al. 2005).

To learn a model we need a set of training data where the ground truth is known, but we cannot assume that we will have training data that is representative of all possible (implicit) queries. Hence, the model should be such that it generalizes to new queries. For that purpose we assume that there is a link between relevance or interest and eye movements, and that this link is, to a reasonable extent at least, independent of the actual topic and query. Then a model of the link can be learned from training data about a subset of possible queries, and it will generalize to new ones. In this paper we formulate such a model and test empirically whether the assumption about a universal link between interestingness and eye movements is useful.

We study whether there is information about interests in the eye movements, and whether it can be extracted by models that make the above-mentioned assumptions. Furthermore, we study whether the usefulness of the inferred interests varies as a function of the amount of system's prior knowledge of the target user topic; while for a new query there is no information available about the potential interestingness of new documents, for a topic already studied for some time there is additional information to be leveraged as well. Namely, the proportion of relevant documents among the set of read documents is higher than in the unread ones, due to the search process so far, and this bias can be utilized in making a content-based proactive search.

As another case study, we investigate whether combining eye movements with document features would help in the standard IR task where explicit relevance feedback is available for a subset of the documents.

2 Earlier work

There has been a lot of interest in implicit relevance feedback techniques in the information retrieval community because they may complement or replace explicit feedback, and thus decrease the need to burden the user. Several implicit feedback measures have been studied before. The literature review by Kelly and Teevan (2003) shows that implicit feedback can indeed improve IR accuracy but there is no consensus about which implicit measures are the most effective.

Claypool et al. (2001) examined how well time spent on a web page, and mouse and keyboard activity can substitute for explicit feedback. They found out that while mouse movements and number of clicks do not correlate strongly with the relevance, reading time and amount of scrolling are good indicators for relevance. Fox et al. (2005) carried out a similar comparison of implicit and explicit feedback. They considered not only the display time and scrolling activity but also other observable measures of user behaviour, such as if the document was printed or bookmarked and how the user exited the page. They observed that the two most important features were display time and the way the user left the page. Although several studies have found the display time to be a good indicator, it may be difficult to analyze in practice. In non-controlled settings the distribution of display times is usually skewed towards zero with numerous outliers (Rafter and Smyth 2001).

Click-through data is another well-studied implicit signal. Joachims et al. (2005) evaluated the quality of click-through measurements on a result page of a web search engine using eye movement measurements (but did not use the eye movements as implicit feedback). They concluded that clicks on the search result page can be used to infer relative relevance judgments between the search results.

Eye movements depend on the type of the visual task the user is performing, which suggests that it is possible to use them to infer the task automatically. For example, Howard and Crosby (1993) noted that the fixations tend to be located sequentially when reading relevant bibliographic citations but non-relevant material is examined non-linearly. King (2002) found out that the distributions of fixation locations differ between reading and counting arrows. She trained a neural network that could separate the two tasks quite efficiently using just the eye movements.

Use of eye movements in IR is a relatively new approach. Maglio et al. (2000) and Maglio and Campbell (2003) introduced a prototype attentive agent application which monitors eye movements while the user views web pages, in order to determine whether the user is reading or just browsing. If reading is detected, more information of the topic is sought and displayed. The feasibility of the application was not, however, experimentally verified.

Eye movements have been used in applications that could be broadly classified as eye-movement-based user interfaces, one of the most known examples being a fast predictive eye typing system Dasher (Ward and MacKay 2002). More recently, Fono and Vertegaal (2005) introduced EyeWindows, an attentive windowing technique which uses eye tracking, rather than manual pointing, for selecting focus windows.

Eye movements were first used in an information retrieval task by Salojärvi et al. (2003, 2005a). Discriminative hidden Markov models were applied to estimate the relevance of lines of read text, and the performance of the method was verified in a

controlled experiment. A competition was subsequently set up, where the participants competed in predicting relevance based on the eye movements (Puolamäki and Kaski 2006).

A prototype information retrieval system was introduced by Puolamäki et al. (2005). The system used relevance information combined with collaborative filtering to seed out relevant scientific articles, the task being to infer if the user found text snippets relevant or not. This earlier prototype did not use the textual content of the documents at all, and hence it could not be used to predict the relevance of unseen documents without some other source of information, such as collaborative filtering.

The methods used so far can be characterized as attempting to use eye movements to assess directly the relevance of displayed information. The methods vary from simple evaluation of attention as in EyeWindows to prediction of relevance from the type of eye movements involved, as in the example of discriminative hidden Markov models. In the current paper we use eye movements in a meta-learning task to infer a weighting over terms that can be used to predict the relevance of unseen documents for the user. This can be seen as an extension of earlier work, in the sense that during the application phase the eye movements are processed by the learned function rather than used directly as features for a retrieval algorithm. This approach opens up the possibility of tapping a rich source of potential information about the user through implicit inferences made without the need for direct querying or input.

The experimental setup used here was introduced first in a conference publication containing a brief feasibility study (Hardoon et al. 2007). That was the first study where documents have been ranked based on their textual content and eye movements of the user. Now we extend and complete the previous study, in particular by addressing the issue of bias in the proportion of relevant documents, which is important for a realistic information retrieval scenario where the user is likely to see varying proportions of relevant documents, and by studying the importance of the features for the task. We also provide a detailed analysis and discussion of the methods and related work.

3 Problem and approach

Our objective is to predict from term-specific eye movements a query vector that can be used to evaluate the relevance of yet unseen documents. Our approach is explained in Sect. 3.1. We also utilize the same overall framework in the case where some explicit relevance feedback is available, and we combine explicit feedback with eye movements. This latter task is discussed in Sect. 3.2.6.

3.1 Overall algorithm

We work with the bag-of-words (BOW) representation of the documents. The goal is to construct a query function $g_{\mathbf{w}}(\mathbf{d})$, where $\mathbf{d}$ is a BOW representation of a new document, and g is a two-class classifier which predicts whether $\mathbf{d}$ is relevant or not. The parameter vector $\mathbf{w}$ represents the implicit query; our task is to provide the classifier with such a $\mathbf{w}$ that it will classify well according to the user's interests. The function g is parametrized such that there is a specific parameter w_t for each term t.

We assume that there is a link between the eye movements and the importance of a word for the query. More specifically, we assume a parametric functional form $w_t = f_{\boldsymbol{\lambda}}(\mathbf{e}_t, \mathbf{s}_t)$ for the relationship between eye movement features collected during reading, denoted collectively by $\mathbf{e}_t$ for term t, query-independent parameters $\mathbf{s}_t$ associated with the term t (e.g. inverse document frequency of term t) and the query parameters w_t. The eye movement features describe the way in which the term is viewed (for example, fixation duration on the term and saccade lengths before and after viewing the term). The features are described in more detail in Sect. 3.2.4. The $f_{\boldsymbol{\lambda}}$ could in principle be any predictor, with its parameters specified by $\boldsymbol{\lambda}$.

The parameters $\boldsymbol{\lambda}$ of the predictor $f_{\boldsymbol{\lambda}}$ are learned on a set of training tasks, where the true interest of the user is known. Following our assumptions laid out above, the functional form of the predictor $f_{\boldsymbol{\lambda}}$ is topic-independent. Hence, the training step needs to be done only once. After that the predictor can be applied to previously unseen queries to produce the query parameters $\mathbf{w}$.

In this work we choose the query function $g_{\mathbf{w}}(\mathbf{d})$ to be a Support Vector Machine (SVM) with the parameter vector $\mathbf{w}$ formed of the term-specific parameters w_t. As a predictor $f_{\boldsymbol{\lambda}}$, which gives the parameters of the SVM, we use standard linear and non-linear regressors (details later). We purposely use standard state-of-the-art machine learning methodologies in this proof-of-concept work to make the approach easily expandable.

3.2 Training

Our main task is to formulate an IR query, using the eye movements as the only feedback signal. The query need not, however, be understandable by humans; in fact, it suffices to formulate the query in such a way that it can be used by the query function $g_{\mathbf{w}}(\mathbf{d})$ to predict relevance for new documents $\mathbf{d}$.

The training data consists of a set of documents and the eye movements of persons who read them while they were searching for documents of certain known topics. Our aim is to try to infer a query from the eye movement recordings and BOW vectors of the read documents, such that the relevance of new, unseen documents to the topic can be predicted. More detailed description of the data is provided in Sect. 4.1.

The training consists of two phases: first, we learn the predictor parameters $\boldsymbol{\lambda}$ by optimizing $f_{\boldsymbol{\lambda}}$ to produce query vectors that are good at separating the known topics in the training set. Next, we apply the learned predictor to a previously unseen topic to infer a query vector $\mathbf{w}$. Finally, we evaluate the performance by ranking unseen test documents according to the query function $g_{\mathbf{w}}$.

The predictor learning phase requires that ground-truth query vectors are known for each topic. These ground truth vectors $\mathbf{w}$ should be such that the query function $g_{\mathbf{w}}$ is able to optimally discriminate between the topics. We approximate the ground-truth by the weight vectors of SVMs that are trained to discriminate the documents of one topic from the others. All SVMs mentioned use the default setting of $C = 1$ and a linear kernel. We call these vectors *ideal weights* and denote the ideal vector for topic c by $\mathbf{w}^c_{\text{ideal}}$. Section 3.2.2 gives more details about the computation of the ideal weights.

The parameter $\boldsymbol{\lambda}$ is optimized by minimizing the squared error

$$\sum_j \left| f_{\boldsymbol{\lambda}} \left(\mathbf{e}_{t(j)}, \mathbf{s}_{t(j)} \right) - w^{c(j)}_{\text{ideal},t(j)} \right|^2 ,$$

where the index j goes over all viewed terms in all training documents, $t(j)$ is the identity of the jth viewed term, and $c(j)$ is the topic of the document where the jth term appeared. After $\boldsymbol{\lambda}$ has been learned, it will be fixed and used in the subsequent steps. We will have either a standard linear least squares regressor or a non-linear sparse-KPLS as the regressor $f_{\boldsymbol{\lambda}}$. They are discussed in more detail in Sect. 3.2.3.

The training set has been constructed to consist of several topics, so that the learned predictor needs to generalize over them and hence become topic-independent. We will test whether this is the case by evaluating the performance on a topic that was not a part of the training set.

The second step in inferring the query for a new topic is constructing the vector $\mathbf{w}$ by letting $w_t = f_{\boldsymbol{\lambda}}(\mathbf{e}_t, \mathbf{s}_t)$ for all terms t that were viewed during the test query. If a term t appears in multiple documents or multiple times in one document, the corresponding feature vector $(\mathbf{e}_t, \mathbf{s}_t)$ will be set to the average over all the occurrences. Zero is assigned to non-viewed terms. After forming the $\mathbf{w}$ we use the query function $g_{\mathbf{w}}$ to classify test documents.

3.2.1 Cross validation methodology

We have been specifically careful in designing the experiments such that testing data never affects the learning, and that is why the following procedure may appear slightly complicated. We evaluate the performance using a cross-validation approach where we leave out one topic at a time for testing and use the rest for training. We recorded eye movement data from several test subjects while they were reading text snippets, looking for documents about a given topic. We had 25 such search topics in total. In addition to the documents that were shown during the experiments we have a separate test set of documents without eye movement recordings for evaluation purposes.

We want to learn a mapping, from the eye movements to the weights, that is independent of the topic. The eye movement features from the testing topic need to be excluded from the training. To this end, the features are partitioned into two sets. The first set, T_1, includes the features from the training topics (other topics besides the left-out topic). More specifically, T_1 is the collection of $(\mathbf{e}_t, \mathbf{s}_t, c)$ triplets for all words viewed during the training topics, where $\mathbf{e}_t$ and $\mathbf{s}_t$ are the eye movement and query-independent features of term t, and c is the search topic that the user was looking for when this sample was generated. The second set, T_2, includes feature pairs $(\mathbf{e}_t, \mathbf{s}_t)$ for the viewed words in the left-out topic.

T_1 is used for training the parameter vector $\boldsymbol{\lambda}$, and T_2 for inferring the query vector for the left-out topic. Finally, the learned query is evaluated on an independent test set. Because the set T_1 does not contain documents from the left-out topic, the predictor cannot specialize on the left-out topic. Instead, if it is able to learn to perform well on the left-out topic, it must be independent of the query and in that sense universal. The approach is summarized in the pseudocode in Algorithm 1.

Algorithm 1 Pseudocode of the training and testing procedure

// 1. Construct the ideal weights
for each topic c
    Train an SVM to discriminate BOW vectors between topic c and other topics
    Let $\mathbf{w}^c_{ideal}$ be the parameter vector of the SVM

for each topic c' // c' is the left-out topic
    // 2. Training

    // Learn the regressor parameters $\boldsymbol{\lambda}$
    $T_1 \leftarrow$ a set $\{(\mathbf{e}_t, \mathbf{s}_t, c)\}$ of all viewed words t in the training set, where topic $c \neq c'$
    $\boldsymbol{\lambda} \leftarrow \arg\min_{\boldsymbol{\lambda}} \sum_{j \in T_1} |f_{\boldsymbol{\lambda}}(\mathbf{e}_{t(j)}, \mathbf{s}_{t(j)}) - w^{c(j)}_{ideal,t(j)}|^2$

    // Compute the query vector $\mathbf{w}$ for the left-out topic c'
    $T_2 \leftarrow$ a set $\{(\mathbf{e}_t, \mathbf{s}_t)\}$ of all viewed words t in the left-out topic c' (average if t occurs several times)
    $w_t \leftarrow f_{\boldsymbol{\lambda}}(\mathbf{e}_t, \mathbf{s}_t)$ for all samples in T_2

    // 3. Testing
    Rank the unseen test documents $\mathbf{d}$ from a separate test set according to $g_{\mathbf{w}}(\mathbf{d})$
    Compute MAP (c' is the positive topic and all others are negative)

3.2.2 Ideal weights

We use as ground-truth the weight vector of an SVM which has learned to predict, based on full knowledge of the topic and content of learning documents, whether or not a document belongs to the given topic. We call these ground-truth values *ideal weights* and use them as targets when training a regressor $f_{\boldsymbol{\lambda}}$ for the same topic. In principle, one could select a different classifier to construct the ideal weights but for ease of integration with the remainder of our system we opt for SVM. The input for this ideal weight SVM is the BOW representation of the document, and the label is $+1$ if the document is categorized to the current topic, or -1 if it is not. These labels are known for the training documents. Because the ideal weights are computed using one SVM per search topic, we get one weight value for each term in the dictionary for each search topic; the weight represents the term's "fit" to the given search topic.

We will use the ideal weights to train a regressor described later in Sect. 3.2.3. The regressor is then used to construct a new classifier for predicting the relevance of the unseen documents.

3.2.3 Regression

We use two types of regressors for predicting the terms' weights w_t from the measured eye movement and query-independent textual features, $\mathbf{e}_t$ and $\mathbf{s}_t$. The simplest mapping we employ is a standard linear regressor

$$f_{\boldsymbol{\lambda}}(\mathbf{e}_t, \mathbf{s}_t) = \sum_i \lambda^e_i e_{ti} + \sum_i \lambda^s_i s_{ti},$$

where the parameter vector $\boldsymbol{\lambda}$ is divided into two parts, $\boldsymbol{\lambda}^e$ and $\boldsymbol{\lambda}^s$. The first one contains the regression coefficients related to the eye movement features $\mathbf{e}$ and the latter the coefficients related to the query-independent features $\mathbf{s}$.

In addition to least squares regression, we use a non-linear sparse dual Partial Least Squares (PLS) approach (Dhanjal et al. 2006). This method uses a general framework for feature extraction based on a kernel PLS (Rosipal and Trejo 2001) deflation method. KPLS maps the feature vectors nonlinearly to a feature space, and does ordinary least squares estimation in the new space. The sparse dual PLS selects, in each iteration, the projection for the least squares regression to be a subset of samples that have maximal covariance with the label. In other words, in each iteration of the algorithm a subset of the kernel is computed and the sample (with the feature combination) that gives maximal covariance is selected as the projection.

In the PLS framework we still adhere to our prior assumption of learning a mapping from eye movements to ideal weights, whereas the eye movements are now kernelized. Due to the large number of samples (eye movements) we are unable to compute the full kernel matrix and therefore only compute a small random portion of the kernel at each iteration of the sparse-KPLS algorithm. For a detailed account of sparse-KPLS we refer the reader to (Dhanjal et al. 2006).

We use a Gaussian kernel with the sparse-KPLS where the width parameter σ for the kernels is optimized, per search topic, using tenfold cross validation on the training data.

We normalize the query vector $\mathbf{w}$ in the 2-norm.

3.2.4 Features

The gaze direction is an (indirect) indicator of the focus of attention, since accurate viewing is possible only in the central fovea area (1–2 degrees of visual angle). The correspondence is not one-to-one, however, since the attention can be shifted without moving the eyes. The eye movement trajectory is traditionally divided into fixations, during which the eye is fairly motionless, and saccades, rapid eye movements from one fixation to another.

The fixations during reading have previously been observed to last 200–250 ms on average and rarely less than 100 ms (Levy-Schoen and O'Regan 1979). Furthermore, Granaas et al. (1984) have showed that moving text two letters positions at a time at 88 ms intervals hinders reading comprehension substantially, which indicates that sub 100 ms durations are too short for effective reading.

We identified fixation locations from the measured eye movement trajectories by windowing; if the successive points stayed inside 30 pixel square (about 0.6 visual angle for a person sitting at 60 cm distance from the screen) for more than 100 ms, they were considered to form one fixation. Every fixation was mapped to the closest word, unless the fixation occurred well outside any text (at least 1.5 times the text height), in which case it was discarded. In this paper we only report the results for the fixation time cutoff of 100 ms. However, we ran the same analyzes with the fixation time cutoff of 40 ms, which is the value recommended by the eye tracker manual. The results with the 40 ms cutoff were very similar to the 100 ms cutoff reported in this work.

We extracted 22 eye movement features (denoted by $\mathbf{e}_t$) from the recorded eye movement data for each term, and 4 text features (denoted by $\mathbf{s}_t$) for the target words of the fixations. The features are listed in Table 1. The eye movement features we are

Table 1 List of features

Eye movement features		
1	Integer	Number of fixations to the word
2	Integer	Number of fixations to the word when the word is first encountered
3	Binary	Did a fixation occur when the line that the word was in was encountered for the first time?
4	Binary	Did a fixation occur when the line that the word was in was encountered for the second time?
5	Continuous	Duration of the previous fixation when the word was first encountered
6	Continuous	The duration of the first fixation when the word was first encountered
7	Continuous	Sum of durations of fixations to a word when it is first encountered
8	Continuous	Duration of the next fixation when the gaze initially moves on from the word
9	Integer	Distance (in pixels) between the first fixation on the word and previous fixation
10	Integer	Distance (in pixels) between the last fixation on the word and the next fixation
11	Integer	Distance (in pixels) between the fixation preceding the first fixation on a word and the beginning of the word
12	Integer	Distance (in pixels) of the first fixation on the word from the beginning of the word
13	Integer	Distance (in pixels) between the last fixation before leaving the word and the beginning of the word
14	Continuous	Sum of all durations of fixations to the word
15	Continuous	Mean durations of fixations to the word
16	Integer	Number of regressions leaving from the word
17	Continuous	Sum of durations of fixations during regression initiating from the word
18	Binary	Did a regression initiate from the following word?
19	Continuous	Sum of the durations of the fixations on the word during a regression
20	Continuous	Mean pupil diameter during fixation
21	Continuous	First fixation duration divided by total duration of fixations on the display
22	Integer	Number of words skipped since previous fixation
Textual features		
23	Integer	Length of the word
24	Continuous	Position of the word in the document divided by total number of words in the document
25	Continuous	Position of the word in the line divided by the line length
26	Continuous	Logarithm of inverse document frequency of the word

using have been previously described in a technical report (Salojärvi et al. 2005b). These are typical features in psychological studies (Rayner 1998).

We also used four query-independent features which do not depend on the way the user viewed the document but only on its textual content. The length of the word and the inverse document frequency are related to the level of mental processing required to comprehend the word. The relative position in the document may be related to the relevance of the word, if, for example, the user usually reads only the beginning of the document but sometimes also more if he finds the topic is interesting.

3.2.5 Explicit feedback SVM model

If explicit relevance feedback, that is, the relevance of the training documents, was available and we knew that the implicit query remained the same in the test documents, we would not need to infer the query vector from the eye movements. Instead, it could be computed more directly from the BOW vectors of the training documents. We could simply train an SVM to classify between the training topic and other topics.

Compared to the approach of the previous section this is much simpler. We can skip computing the ideal weights and learning the regressor $f_{\boldsymbol{\lambda}}$ and, instead, learn the query function $g_{\mathbf{w}}(\mathbf{d})$ more directly. We let the query function to be an SVM that discriminates between the relevant and non-relevant documents in the current topic. The parameter vector $\mathbf{w}$ is learned during the optimization of the SVM.

The inputs to the SVM are document BOW vectors, similarly to the SVMs that were used in computing the ideal weights (see Sect. 3.2.2). The differences are that the training labels are now the true relevance labels instead of the category labels, and that the training set consists of only the documents shown during the searches where the user's true interest was the current topic.

The learnt weight vectors are used to classify the test documents. The resulting classifier is referred to as SVM_{ex}. It represents the "best imaginable" performance, which implicit feedback cannot realistically be expected to outperform.

3.2.6 Combining eye movements and explicit feedback

Next we consider a situation where we observe both the true relevance labels and eye movements. We want to find out whether using both of these information sources in conjunction to predict the query could improve the classification accuracy, in comparison with the model of the previous subsection that uses only explicit feedback.

Because we again have the explicit relevancy for the documents we can skip the learning of ideal weights and the regressor $f_{\boldsymbol{\lambda}}$, and learn the query vector w directly. The query function $g_{\mathbf{w}}(\mathbf{d})$ predicts the relevancy of a document given both the textual content and the eye movements on the document. We choose g to be a two-view classifier called SVM-2K.

We regard the textual content as one representation and the measured eye movement feature vectors as a second representation of the document. We project the two

views through distinct feature projection, thus creating two kernels, one for each representation.

The Kernel Canonical Correlation Analysis (KCCA, Hardoon et al. 2004) algorithm looks for directions in the two feature spaces such that when the training data is projected onto those directions the two vectors (one for each view) of values obtained are maximally correlated.

A straightforward way to do classification using two data sets would be to first project the data into the KCCA space and then train an SVM classifier. Though this sequential approach seems effective (Meng et al. 2005), there appears to be no guarantee that the directions identified by KCCA will be best suited to the classification task. Farquhar et al. (2006) have shown that the two distinct stages of KCCA and SVM can be combined into a single optimization that was termed as SVM-2K. The training of an SVM-2K model is explained in Appendix A.

3.3 Testing

After the query vector has been learned in the training phase we can predict the relevance of unseen test documents. We rank the test documents according to the values of the discriminant function $g_{\mathbf{w}}(\mathbf{d})$. The discriminant functions for the implicit, explicit, and combined feedback tasks are described in the following subsections.

If the classifier is accurate the test documents belonging to the current left-out category should be near the top of the resulting list. The quality of the ranking is measured by average precision (see Sect. 4.3).

3.3.1 Implicit feedback regression models

The compatibility of a test document BOW vector $\mathbf{d}$ to the inferred query $\mathbf{w}$ is computed by an SVM discriminant function. For linear kernel the function is

$$g_{\mathbf{w}}(\mathbf{d}) = \mathbf{w}^T \mathbf{d} + b,$$

where b is the bias that is learned together with the weights $\mathbf{w}$. The vectors $\mathbf{w}$ and $\mathbf{d}$ are normalized in 2-norm to keep the length of the document from affecting the similarity measure.

Instead of thresholding the values of the discriminant function, as would be done in a pure classification task, we rank the test documents according to values of the function. As we are only interested in the ranking, the bias term can be left out from g. We measure the goodness of the ranking by average precision (see Sect. 4.3).

3.3.2 Explicit feedback SVM model

The discriminant function is of the same form as in the case of implicit feedback. Of course, the weight vector $\mathbf{w}$ is now computed using the explicit feedback.

3.3.3 SVM-2K combining implicit and explicit feedback

The discriminant values of the SVM-2K model for a test document are given by the (1) in Appendix A. Because no eye movement measurements are available for the test documents, the eye movement feature projection ϕ_A in (1) is set to zero when computing the decision values for the test documents.

4 Experiments

4.1 Data collection

4.1.1 Document corpus

The document set consists of 750 documents from the Wikipedia. Of them, 500 were used to form the training set and the rest were the test set. The documents have been partitioned into 25 categories by the editors of the Wikipedia. There were 12–23 documents from each topic in the training set and ten documents per topic in the test set. The categories and their sizes are listed in Table 2.

The documents were represented as Term Frequency-Inverse Document Frequency (TFIDF, Salton and McGill 1983) vectors. If the document corpus is denoted by D and the dictionary by T, then the TFIDF weight for a term $t \in T$ in a document $i \in D$ is the product of term frequency in the document and logarithm of the inverse document frequency of the term in the corpus:

$$\mathrm{TFIDF}(i,t) = n_{it} \log \frac{|D|}{\left|\left\{j \in D \mid n_{jt} > 0\right\}\right|},$$

where n_{it} denotes the number of occurrences of term t in document i. The dictionary T consisted of all stemmed words in the documents except for numbers and some frequent 'stop' words like 'of' and 'the', which were omitted. Also, words that only appeared in single interest categories were removed from the data in order to create a more realistic scenario, i.e. overlapping words through the different categories. The dictionary included 5306 terms in total.

We truncated the documents so that each had at most 11 lines of text, in order to fit them to the screen. We wanted to keep the test setup as simple as possible, and to avoid the need to scroll the text. To avoid any artifacts that might attract unwanted eye movements, the titles of the documents were removed and no sentence was cut in the middle. We made sure that the content of each truncated document was sufficient for inferring its topic by manually inspecting all documents. We assume that, because the categories are quite diverse and it is fairly easy to categorize the truncated documents into the different categories, the participants would not have needed scrolling in most cases even if it had been possible. This is further supported by the fact that only a minority of the documents had any fixations on the final lines (see Fig. 1).

Table 2 Summary of the training and test corpora

Topic	Number of documents in the training set	Number of documents in the test set
Astronomy	22	10
Ball games	21	10
Cities	12	10
Court systems	23	10
Dinosaurs	17	10
Education	22	10
Elections	21	10
Family	16	10
Film	20	10
Government	20	10
Internet	23	10
Languages	21	10
Literature	21	10
Music	16	10
Natural disasters	19	10
Olympics	19	10
Optical devices	20	10
Postal system	22	10
Printing	22	10
Sculpture	20	10
Space exploration	19	10
Speeches	23	10
Television	22	10
Transportation	22	10
Writing systems	17	10
Total	500	250

4.1.2 Participants

There were ten participants in the experiments. They were voluntary post-graduate and senior researchers from Department of Information and Computer Science, Helsinki University of Technology.

4.1.3 Experimental procedure

We measured users' eye movements when they were reading documents in order to classify the documents into interesting and uninteresting. We artificially constrained users' interest by giving them a topic and asked them to identify documents which were related to that topic. Topics were the Wikipedia categories listed in Table 2.

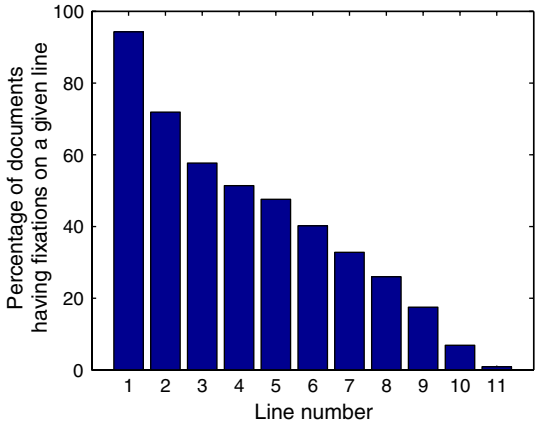

Fig. 1 Proportion of documents having fixations on a given line

Fig. 2 A sample document (on the *left*) and a screen which the user gets when he has finished reading (on the *right*)

The user read ten short documents trying to recognize those that were related to the given topic. He was instructed to read the document until he could say whether or not the document was related to the topic and, after making up his mind, to press any key in the keyboard. The key press then replaced the document by a form where the user reported his impression of the relevance by pressing one of two possible keys. The next document was shown immediately after the user had reported his opinion. A sample document and a feedback screen are shown in Fig. 2.

We call a combination of one topic and the 10 documents the user read while searching for that particular topic a *session*. Each user completed ten sessions, with a different search topic and documents in each. A short break was allowed after the third and the sixth session.

On average half of the training documents in a session were relevant and the rest were randomly drawn from unrelated topics. In total 1,000 documents ($= 10$ users $\times$ 10sessions/user $\times$ 10 documents/session) were displayed during the tests. This means that each of the 500 training documents was shown to two users on average.

The answers given by the users agreed in 97% of the cases with the true labels.

4.1.4 Eye tracker

The gaze locations on the screen were recorded with a Tobii 1750 eye tracker while the test subjects were reading the documents. The Tobii system consists of an infra-red LED and two cameras mounted on the frame of a computer screen. Tobii measures gaze direction 50 times per second by illuminating both eyes with infrared and measuring the light reflected from the cornea. The system is fairly robust to head movements. The user was sitting at a 60 cm distance from a 17 inch computer screen. The text was displayed with quite a large font and spacing so that it would be possible to map the gaze location to the correct word reliably. There was room for at most 11 lines of text on the screen. The eye tracker was calibrated in the beginning of the experiment for each user and after every break.

4.2 Experimental setup

4.2.1 Implicit feedback models

As was discussed in Sect. 3.2.3, we tried both the linear least squares and the non-linear KPLS regressors for predicting the term-specific weights from the eye movement features. The model using linear regressor is referred to as W_{lin}, and the model using KPLS is called $W_{\text{eye+text}}(26)$, where 26 is the dimension of KPLS feature space. Both of these models use the eye movement features $\mathbf{e}_t$ and the query independent textual features $\mathbf{s}_t$. To determine how much the extra information from the eye movements boosts the performance we trained a second model that is similar to $W_{\text{eye+text}}(26)$, but does not use eye movements at all. To be exact, the differences are that the second model uses only the query independent features, and that the features are computed for each of the words appearing in the document, not just for the viewed words. We refer to this second model as $W_{\text{text}}(4)$, where 4 is again the number of projection directions in KPLS. The number of projection dimensions are chosen to coincide with the number of features in the corresponding regressor. We have tested with other numbers of directions, too, and the performance does not change much when the number of directions is increased (data not shown here; see Hardoon et al. 2007).

The above regression scheme is performed by pooling the eye movements from all users who completed the same query. The size of training set used to adapt the model to a particular search topic (set T_2 in Algorithm 1) is 30–50 documents depending on the topic (3–5 test subjects performed the same query and each of them saw ten documents). About half of them were relevant. If the users employ very different reading strategies it might make more sense to handle them separately. Therefore we also test a model that is otherwise identical to $W_{\text{eye+text}}(26)$ but instead of pooling the data each session is handled as a separate, smaller training set consisting of only ten documents. This model is referred to as $W_{\text{us}}(26)$ (for "user specific") in Table 3.

Figure 3 shows a user's eye movement trajectory on a document, and the term weights on the same document inferred by the $W_{\text{eye+text}}(26)$ regressor from eye movements of all test subjects who were searching for documents about *Dinosaurs*.

Table 3 The mean average precisions for biased data

Random (no feedback)	TFIDF (no feedback)	W_{lin} (impl. feedback)	$W_{\text{text}}(4)$ (impl. feedback)	$W_{\text{eye+text}}$ (26) (impl. feedback)	$W_{\text{us}}(26)$ (impl. feedback)	SVM_{ex} (impl. feedback)	SVM-2K (impl. feedback)
6.0	73.4	37.8	19.9	46.3	31.3	75.1	77.0

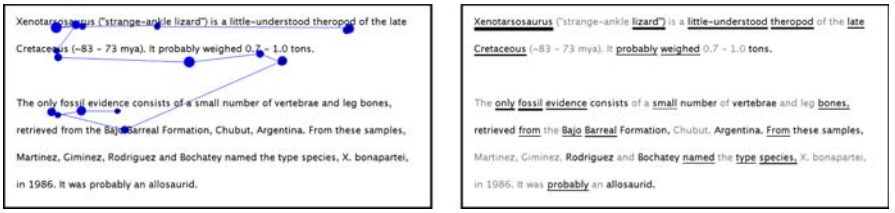

Fig. 3 On the *left*, eye movement trajectory. The document is about *Dinosaurs*, which is also the topic the reader was interested in. The circles mark the fixations, and the radius of a circle is proportional to the duration of the fixation. On the *right*, weights inferred from eye movements. The *thickness* of the *underline* denotes the magnitude of the weight. *Gray* words do not appear in the dictionary

4.2.2 Explicit feedback models

For reference we also tested how much the performance can be improved if explicit relevance feedback is available. The explicit feedback can be used alone or combined with the information from the eye movement measurements. If the eye movements are excluded, only the TFIDF vectors of the documents are available. In this case we used a standard SVM classifier that was trained to discriminate between relevant and non-relevant samples (Sect. 3.2.5). The resulting model is called SVM_{ex}. To combine explicit feedback and eye movements we used the SVM-2K model of Sect. 3.2.6. This model is referred to as SVM-2K below.

4.2.3 Baseline models

We compare the results to two baseline models that do not utilize any relevance feedback at all. The simpler one just ranks the test documents randomly.

The second baseline model ranks the test documents by the average similarity of their textual content with the training documents. The training and testing documents are presented as 2-norm normalized TFIDF vectors. We compute mean cosine distance (which equals to dot product for normalized vectors) of the test documents to the training documents. More formally, for each training set s with documents $\mathbf{y}_j$, $j \in Tr(s)$, and for each test document $\mathbf{x}_i$ we compute

$$g(\mathbf{x}_i) = \frac{1}{|Tr(s)|} \sum_{j \in Tr(s)} \frac{\mathbf{x}_i^T \mathbf{y}_j}{||\mathbf{x}_i|| ||\mathbf{y}_j||},$$

where $Tr(s)$ is the set of all (both positive and negative) training documents in the training set s and the document lengths are taken into account by normalizing the vectors. The test documents are then sorted according to the cosine distance, in the order of similarity with the training documents.

4.2.4 About bias in the learning set

In a typical information retrieval setup, the proportion of relevant documents the user receives during a search session will be larger than on average in the corpus. This happens because an information retrieval system naturally aims for a good precision, or proportion of relevant documents of all results. This creates a bias: the proportion of relevant documents is larger in the learning set (seen documents) than in the test set (unseen documents). If the precision of the search results is good enough, then as a result of the bias even the learning document set as such is a reasonably good query—without any explicit or implicit relevance feedback! This has a direct effect on the performance of any information retrieval algorithm. Hence we will inspect the performance of our methods as a function of the bias.

In our experiments, half of the documents in a session were positive on average and others were randomly sampled from the negative topics. Because the positive documents form a large cluster while negative documents are likely to be more scattered, even a simple TFIDF baseline model can perform well in this setup as the mean vector is drawn near to the large positive cluster. On the other hand, in a more realistic setup we can't assume such a large fraction of the seen documents would be relevant, and therefore models using just the textual content of the documents are likely to perform much worse. Implicit or explicit feedback can help the search engine to identify the relevant documents even when the bias is low.

To test our algorithm under smaller bias we created new training sets by re-balancing the data in the original sessions. We constructed new artificial sessions by dividing the documents (and their corresponding eye movement measurements) anew. The new sessions have only one positive document, and therefore the positive topic will not stand out from the rest. We also limited the proportions of the topics in the test set to be the same as those in the training set. This way there is no extra information available in the setup that could bias the results, and the random ranker is a fair baseline.

Each session was divided into n new sets by taking one of the n positive and all negative documents in the original session. If two or more negative documents happened to belong to the same topic, all possible one-document-per-topic combinations were used as the new training sets constructed from that session. Unlike in the original setup, where we always used all test documents for testing, we now construct a separate test set for each training set by taking only the test documents from the same topics that appear in the training set. Thus, the proportion of the topics in the training and testing is the same but the size of the test set depends slightly on the training set. We call these resampled sets *unbiased* training sets.

There were some documents where the eye tracker had failed to capture any eye movements at all. A likely reason is that the user was probably sitting too far or too close to the monitor while reading these documents. When constructing the unbiased

sessions we left out all sessions which would have included at least one of these documents without any eye movements. In total, we get 694 unbiased sessions which had 88 distinct pairs of user and topic.

Training of the classifiers for the unbiased case is done identically to the training in the biased case in Sect. 3.2. The learning algorithm again iterates over all sessions and learns the regression parameters from data collected from all other topics except the left-out topic. The training set (denoted by T_1 in Sect. 3.2.1) again consists of features collected from all the sessions except for the left-out topic. Because T_1 is the same for all new sessions which were constructed from a single original session, also the regression parameters will be the same for those sessions. The training set T_2 for learning the term weights w_t contains $(\mathbf{e}_t, \mathbf{s}_t)$ feature pairs from the documents in the left-out session. The sessions with the same topic are not pooled, as was done in the biased case, because doing so would alter the amount of bias. The size of T_2 is 4–8 documents depending on the session. The model is learned and the predictions are made the same way as before.

In the testing phase the performance of the learned models is tested on the corresponding test sets. To get a single performance index for each user/topic pair, we computed averages of the performance figures over sessions which had the same combination of users and topics.

We compare the performance of the algorithms on the biased and the unbiased data sets.

4.3 Results

4.3.1 Performance measure

To compare the goodness of the rankings of the unseen test documents produced by the different methods we computed *mean average precisions* (MAP).

Average precision measures how well the positive documents are positioned in a given ranking. It is calculated as a mean of average value of precision at the positive ranks:

$$\text{AVGPREC} = \frac{1}{R} \sum_{i=1}^{R} \frac{i}{r_i},$$

where R is the number of positive examples in the test set and the r_i are the rankings of the positive examples, ordered such that $r_i < r_{i+1}$. The mean average precision is the mean of average precisions of different information needs or, our case, sessions.

We inferred a query $\mathbf{w}$ with each method for each session and ranked the test documents according to their predicted similarity to the inferred query by computing the discriminant function values $g_{\mathbf{w}}(\mathbf{d})$ for each test document $\mathbf{d}$. The discriminant functions were described in Sect. 3.3. The quality of the ranking was established by computing the average precision. The final MAP value is the mean of the average precisions of all sessions. We computed MAP for all methods both in the biased and unbiased setting. The full results for both cases are in Appendix B.

Table 4 The mean average precisions for unbiased data

Random (no feedback)	TFIDF (no feedback)	W_{lin} (impl. feedback)	$W_{\text{text}}(4)$ (impl. feedback)	$W_{\text{eye+text}}$ (26) (impl. feedback)	SVM_{ex} (impl. feedback)	SVM-2K (impl. feedback)
25.7	25.9	28.1	24.8	27.4	80.0	80.0

4.3.2 About statistical tests

The performances of the models reported in Tables 3 and 4 are significantly different when tested with non-parametric ANOVA ($p \ll 0.0001$ (biased case) and $p = 0.018$ (unbiased case) for differences between all baseline and implicit feedback models, Friedman's Test). This indicates that there is always at least one method for which the performance differs significantly from the others. However, since we are not interested in comparing all methods, but only certain pre-defined pairs, it is sufficient to use pairwise tests.

In the biased case it is clear that TFIDF performs well and random performs poorly. Instead of comparing to those, we are interested to see if the eye movement features bring in additional information compared to using just the text features. If yes, we expect that a model which takes eye movement information into account starts to perform better than text-only models when the bias is reduced. Therefore, we test the significance of $W_{\text{eye+text}}(26)$ versus $W_{\text{text}}(4)$. We additionally investigate the need for user-specific models versus pooling all data together, by comparing $W_{\text{us}}(26)$ versus $W_{\text{eye+text}}(26)$.

In the unbiased case, in addition to testing the difference between eye movement and text-only models, we compare the performance of the eye movement models to the random ranking, which is a fair baseline after the bias is removed.

We use Wilcoxon Signed Rank Test to compare the classification performances of the pairs of algorithms ($W_{\text{eye+text}}(26)$ versus $W_{\text{text}}(4)$; and $W_{\text{eye+text}}(26)$ versus $W_{\text{us}}(26)$). We use the non-parametric Wilcoxon Signed Rank Test instead of paired t-test because the average precision values are not normally distributed. Wilcoxon test is known to be less stringent than the t-test. In comparisons with the random model, however, the Wilcoxon Signed Rank Test is not applicable, because the random model does not have a single ranking of the test set documents, but a distribution of rankings. In comparison with the random model (W_{lin} or $W_{\text{eye+text}}(26)$ vs. random) in the unbiased case in Sect. 4.3.5 we used a standard permutation test for which we sampled 100,000 random rankings.

4.3.3 Baseline performance

The performance of the random baseline model in each session was tested by sampling 100,000 random rankings of the test documents and computing average precisions for each. To summarize the performance of the random model in Table 3, we first computed session-wise mean average precisions over the random rankings and finally took average of them. The MAP is 6.0% in the biased case and 25.7% in the unbiased case.

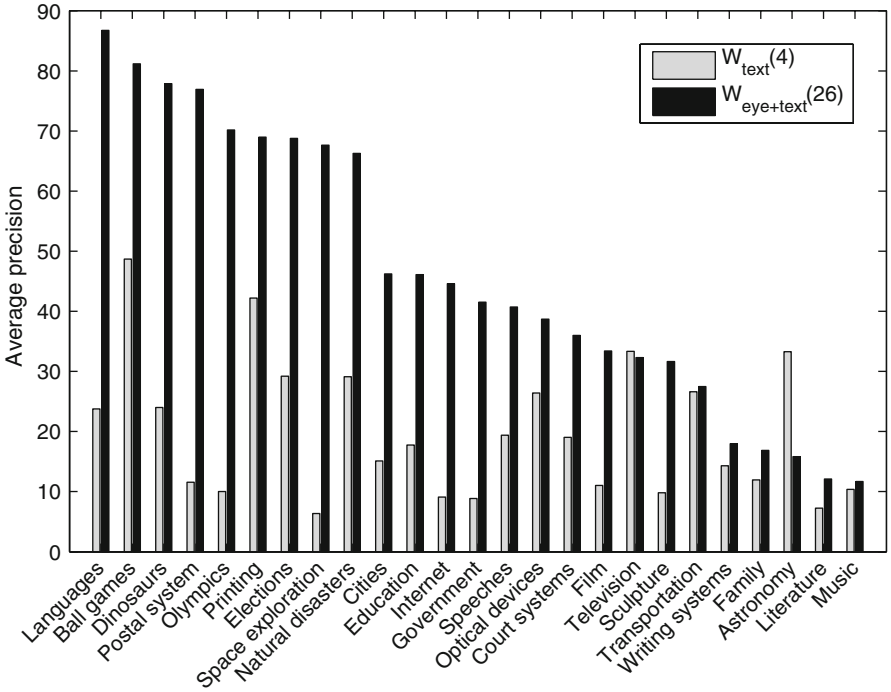

Fig. 4 Average precisions of $W_{\text{text}}(4)$ and $W_{\text{eye+text}}(26)$ in different topics in the biased case. The topics are sorted according to the performance of $W_{\text{eye+text}}$ model

The baseline TFIDF model can often identify the correct topic in the biased setting because the positive documents form the largest cluster among the training documents. This results in achieving a MAP of 73.4%, which is the highest of all tested models in the biased case. In the unbiased case in Table 4, the positive document does not stand out from the rest of the training set, and therefore the performance of the TFIDF model (25.9%) is close to random.

4.3.4 Regression models on biased data

In the biased case (Table 3) the difference in average precision between the non-linear regression model $W_{\text{eye+text}}(26)$, which combines eye movement and textual features, and the textual feature model $W_{\text{text}}(4)$ is statistically significant (46.3 vs. 19.9%, $p = 0.00005$, Wilcoxon Signed Rank Test). This implies that the eye movement features contain information that helps in the complex task of inferring the hidden query. Figure 4 shows the difference in performance of the two models. The availability of eye movement information improves the performance clearly on several topics, but in two cases the text-only model outperforms the eye movement model.

$W_{\text{eye+text}}(26)$, which was trained pooling together the data from all users who completed the same query, is significantly better than $W_{\text{us}}(26)$, which used the user specific average precisions (46.3 vs. 31.3%, $p = 0.0001$, Wilcoxon Signed Rank Test).

This result suggests that the method benefits from the larger training sets created by pooling the data form different users.

We next tested how much the results depend on the specific topics, that is, how universal the results are over the choice of topics. For this, we resampled with replacement topics among the original 25 topics. To study the effect of the number of topics, we repeated the analysis for different set sizes between 2 and 25 topics. For each topic set size, we draw 10,000 random topic sets, and computed the difference in average precision between $W_{\text{text}}(4)$ and $W_{\text{eye+text}}(26)$ in each replicate. We computed BC_a corrected estimates of the 95% confidence intervals (DiCiccio and Efron 1996) for the difference between the two methods based on the obtained bootstrap distribution. To keep the computational workload down we employed the regression parameters we had already learned on the full corpus instead of retraining them on the sampled, smaller topic sets. Note that all training and testing has still been done on completely separate sets. The model using eye movement features was better (the confidence interval included only positive values, i.e. values for which the average precision of the eye movement model was better than that of the text-only model) on all sets larger than three topics. In other words, the model performs reliably regardless of which topics are chosen.

4.3.5 Regression models on unbiased data

The results for the unbiased data are shown in Table 4. The linear regression model W_{lin} attains the highest MAP of all implicit feedback models with the non-linear $W_{\text{eye+text}}(26)$ following closely. Note that the results in Table 4 are not directly comparable to the results on biased data in Table 3, because in the unbiased case the training sets are much smaller (4–8 documents in the unbiased case vs. 30–50 documents in the biased case) and the measurements from different users are not pooled as is discussed in Section 4.2.4.

Combination of eye movement features and textual features gives better performance than the textual features alone ($W_{\text{eye+text}}(26)$ vs. $W_{\text{text}}(4)$, $p = 0.028$). To test our assumption that the implicit feedback can improve over a random ranking when the artificial bias of our training set is removed, we compared the results of $W_{\text{eye+text}}(26)$ to random. As discussed earlier in Sect. 4.3.2 the observed $W_{\text{eye+text}}(26)$ MAP values are compared to the distribution of the MAP value of the random model using a permutation test. We randomly permuted the relevance labels of the test documents 100,000 times in each session, computed the mean average precision, and counted the proportion of the permutations that had at least as high MAP as the eye movement models. Only 2.7% of the permuted samples had at least as high MAP as $W_{\text{eye+text}}(26)$ (27.4% vs. 25.7%, $p = 0.027$, Permutation Test), and 0.26% were at least as good as W_{lin} (28.1 vs. 25.7%, $p = 0.0026$, Permutation Test).

We studied the effect of the choice of topics on the performance using a similar bootstrap approach as in Sect. 4.3.4. We sampled the topics with replacement and constructed sets consisting of 2 to 25 topics. We generated 10,000 samples for each set size and estimated the 95% confidence intervals for the difference in average precision between $W_{\text{text}}(4)$ and $W_{\text{eye+text}}(26)$. The model that uses both text and eye movement features performed significantly better than the text-only model (assessed by the

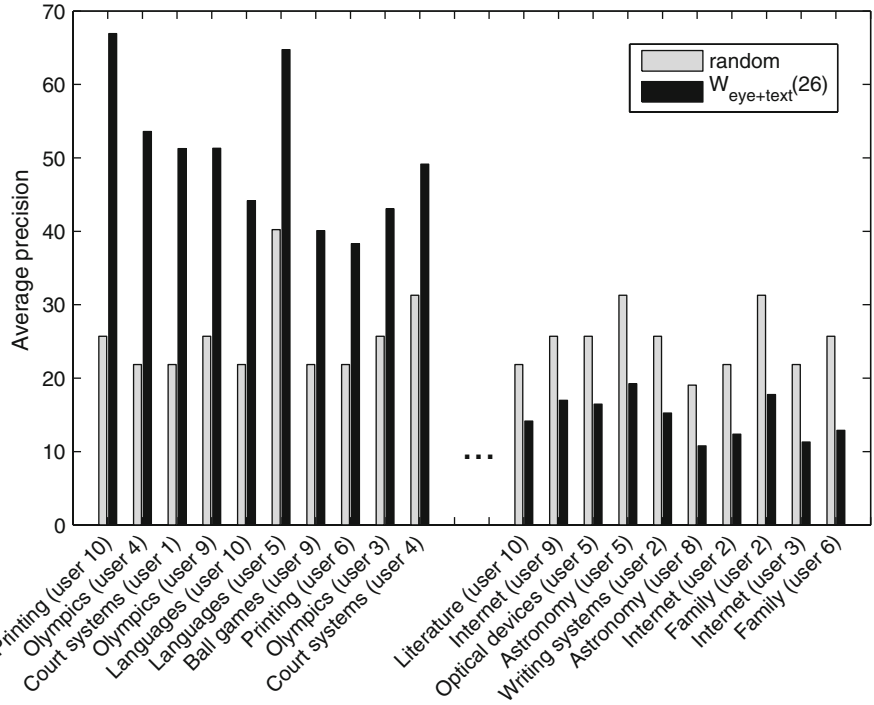

Fig. 5 Average precisions of ten best performing (*left-hand side*) and ten worst performing sessions (*right-hand side*) in the unbiased case. The session are sorted according to the p-value of the Permutation Test between random ranking and $W_{\text{eye+text}}(26)$

confidence interval including only positive values, i.e. values for which the eye movement model was better) with 21 or more topics. The variance of the bootstrap estimate gets larger as the number of topics and thus also the number of average precision samples decreases. At least 21 topics are required for getting statistically significant results with the current amount of training data; it is likely that with more test subjects, or more read documents per topic, fewer topics would be needed for a statistically significant difference.

To get further insight into the performance of the model we took a look at the individual sessions. The average precisions of the best and worst performing sessions (sorted according to the p-value of the Permutation Test between random ranking and $W_{\text{eye+text}}(26)$) are shown in Fig. 5. In 10 out of 88 sessions the average precision of $W_{\text{eye+text}}(26)$ is significantly better than the random model ($p \leq 0.05$, Permutation Test). The average improvement in MAP over random in these queries was 24.5, which is considerably higher than the overall change in MAP for all sessions 1.7 ($= 27.4-25.7$). Some topics appear very frequently among the top-10 sessions; *Olympics* occur three times, and *Court systems*, *Languages*, and *Printing* occur two times each. All of these topics were queried in three sessions in total, except for *Languages* which was the positive topic in four sessions. The fact that they appear so frequently among the highest performing sessions indicates that there

exists queries for which the implicit feedback works well. On the other hand, because the overall MAP of the implicit feedback model is so close to the MAP of random ranking there clearly are other queries which do not benefit from the eye movement feedback.

4.3.6 Feature selection

The eye movement features (Sect. 3.2.4) used in the experiments have been previously proposed in the context of various psychological studies. There is no evidence that all of them are needed for the current task of predicting the interest of a user. It is an important open question what kind of features are best suited for this task.

To get some insight on the relative importance of the features, we analyzed the linear regression model W_{lin} for which such analysis is straightforward. We used the t-test to find out which regression coefficients differed significantly from zero. There were three eye movement features and one text feature that had Bonferroni corrected p-values less than 0.05 in all sessions, both in the biased and unbiased case. These features were "saccade length before first fixation to the word," "did a regression initiate from the following word," "duration of the first fixation to the word divided by the total fixation durations on the document," and "relative position of the word on the document." In addition to these, two features were significant in a subset of sessions either in the biased or the unbiased setting. The feature "duration of the next fixation after leaving the word" was significant in 9 out of 25 sessions in the biased case and 25 out of 694 sessions in the unbiased case, and "duration of a regression starting from this word" was significant in 2 out of 25 biased sessions.

The performance of the linear regression model in the biased setting increased from 37.8 to 40.9% when only the five features with significant contribution in at least nine sessions were selected. On the other hand, the same subset is not optimal for the non-linear regression model. The MAP of $W_{\text{eye+text}}(26)$ decreases from 46.3 to 39.9% with the set of five selected features. In the unbiased case, switching from the full feature set to the set of five important features actually impairs the linear regression model. The MAP decreases from 28.1 to 26.8%. The MAP of $W_{\text{eye+text}}(26)$ increases a little bit from 27.4 to 27.7%. In summary, the extracted features are clearly significant but probably not completely sufficient.

4.3.7 Performance of eye movements combined with explicit feedback

The average precision of the explicit feedback classifier SVM_{ex}, which was discussed in Sect. 3.2.5, is significantly higher than any of the eye movement models. This is, of course, to be expected because explicit feedback gives much more accurate information than eye movements. As a side note, the fact that the TFIDF baseline model attains almost as high performance without any feedback as the SVM_{ex} with known relevance labels shows how substantial the effect of the bias really is in our setup.

Our initial assumption was that combining eye movements with the explicit relevance feedback improves overall performance over using explicit feedback only.

Comparing SVM_{ex} and SVM-2K results for the biased setting in Table 5 shows that this is not true for all search topics. Nevertheless, Table 3 shows that the overall precision for the biased setting is improved slightly by combining the two sources of information. The situation is even more clear in the unbiased setting (Table 4), where the SVM-2K and SVM_{ex} models have equal MAP. The equality of performance is consistent across all categories in Table 6 leading us to believe that the selection of eye movement features used in this study does not improve the overall performance when combined with textual information using SVM-2K. It is apparent that the explicit feedback is sufficient to learn the discrimination in the unbiased case.

5 Discussion

We addressed the extremely hard task of constructing a query in an information retrieval task, given neither an explicit query nor explicit relevance feedback. Only eye movement measurements for a small set of viewed snippets, and the text content of the snippets were available. This is a prototype of a task where the intent or interests of the user are inferred from implicit feedback signals, and used to anticipate the users' actions.

We presented a proof-of-concept solution for this new problem based on the assumption that there is a link between eye movements and the relevance, and that the link is independent of the query. We were able to learn a "universal predictor of relevance predictors" from a collected database of queries, their relevant and irrelevant documents, and the corresponding eye movements. "Universal" here means independent of the actual query in the domain of Wikipedia. We validated the proposed solution on a real dataset, and showed that the predictions were better than those of a simple model which utilized only the textual content of the documents for new queries.

There may be a dependence between the gaze pattern and the topic of interest, even within our set of Wikipedia topics. A topic-specific predictor could therefore perform better for that particular topic, but it would be unable to predict the relevance of yet unseen topics. Our goal was to make universal, or topic-independent, predictor of relevance. We accomplished this by explicitly constructing our predictor such that it is invariant with respect to the permutation of the words—that is, the predictor does not take the semantic meaning of the words into account. The fact that we were still able to reach a statistically significant average prediction results on topics unseen in the training data shows that there is a link between the gaze patterns and interests. The statistically significant features, discussed in Sect. 4.3.6, give hints of the nature of this link. The exact nature or interpretation of this connection needs to be left as a topic of further study.

Notice that any method used to discriminate between categories, based on the BOW representation of the documents, such as ours, requires the categories to have different term frequency distributions. If the term frequency distributions of the categories are too similar to each other the BOW methods are expected to perform worse. In our experiments we used Wikipedia categories which have a relatively good separation in BOW representations. However, it remains an open question whether the

eye movements could effectively be used to find finer distinctions, that is, to find most discriminative words even if the relevant and irrelevant categories were very similar to each other.

We also addressed the issue of 'bias', which is likely to occur in a real world information retrieval situation, where the user is likely to view proportionally more relevant than irrelevant documents. More specifically, when browsing to look for interesting documents, there is no bias, and the bias increases as a function of the length of a focused search or browsing session. Both extremes are important for a practical proactive information retrieval system.

The bias complicates the evaluation of eye tracking results. We showed that taking the eye movement measurements into account improves the precision of the information retrieval, both in the presence and absence of the bias. Eye tracking is more important when there is no bias, since then content-based searches do not help at all. With heavy bias the relative improvement is much smaller. In percentage units the improvements are rather modest, as eye movements are a quite noisy indirect indicator of relevance. Nevertheless, assuming eye tracking signal is cheaply available, it would be a useful source to complement alternative implicit feedback sources. This holds in particular for applications with little bias towards read documents being relevant, such as in browsing applications.

We further experimented with a model where the textual content of the documents and explicit relevance feedback given by the user (whether or not the user thinks the document is relevant to the search topic) were assumed available. As expected, the pure explicit feedback improved the precision significantly to 75.1%. Our results show that also in the biased case, taking the eye movements into account we can further improve the precision by about two of percentage units.

In this paper, we studied whether eye movements are at all suitable as a source of implicit feedback for learning an universal predictor of relevance. The method is still a long way from being usable in practical applications, and there are several important extensions to be studied in the future. One interesting research topic is replacing the general-purpose machine learning components we used here with ones that are better tailored to the task. There are also problems on the practical level: in order to reliably connect gaze location to the correct word using current eye tracker technology, the size of the font has to be fairly large, which restricts the amount of text on the screen. Another factor that would be worth studying is what kind of eye movement features are optimal for the task.

We conclude that in constructing a query, eye movements provide a useful implicit feedback channel. As expected, the feedback obtained from eye movements is less informative than relevance feedback typed in by the user, but nonetheless this implicit feedback can be exploited. The eye movements may be the only available source of relevance information in the unbiased case where nothing can be deduced from the textual content. In practical applications all available feedback channels, in addition to the eye movements, should of course be utilized; the practical implication of this study is the finding that it is a good idea to include eye movement data if they are cheaply available. The future challenge is to improve the gain in precision obtained from the eye movements, in addition to making this new feedback channel more practically applicable.

Acknowledgments AA and SK belong to Adaptive Informatics Research Centre, a centre of excellence of the Academy of Finland. The authors thank Wray Buntine from the Complex Systems Computation Group at University of Helsinki for providing the Wikipedia data, and Craig Saunders and Steve Gunn from the ISIS Research Group at University of Southampton for many fruitful discussions. This work was supported in part by the IST Programme of the European Community, under the PASCAL 2 Network of Excellence, and in part by TKK MIDE programme, project UI-ART. This publication only reflects the authors' views.

Appendix

A Optimization of SVM-2K

SVM-2K combines KCCA-projection and a regular SVM by introducing the constraint of similarity between two 1-dimensional projections. The extra constraint is chosen slightly differently from the 2-norm that characterizes KCCA, which typically finds a (varying in number) sequence of projection directions that then can be used as the feature space for training an SVM. Denote the two projections (views) of a training sample x_i by $\phi_A(x_i)$ and $\phi_B(x_i)$. In order to obtain sparse set of projection directions we take an ϵ-insensitive 1-norm using slack variables to measure the amount by which points fail to meet ϵ similarity:

$$|\langle w_A, \phi_A(x_i)\rangle + b_A - \langle w_B, \phi_B(x_i)\rangle - b_B| \leq \eta_i + \epsilon,$$

where w_A, b_A (w_B, b_B) are the weight and threshold of the first (second) SVM, and η_i are slack variables. Combining this constraint with the usual 1-norm SVM constraints and allowing different regularization constants gives the following optimization problem:

$$\begin{aligned}
\min \quad & L = \tfrac{1}{2}\|w_A\|^2 + \tfrac{1}{2}\|w_B\|^2 + C^A \sum_{i=1}^{\ell} \xi_i^A + C^B \sum_{i=1}^{\ell} \xi_i^B + D \sum_{i=1}^{\ell} \eta_i \\
\text{such that} \quad & |\langle w_A, \phi_A(x_i)\rangle + b_A - \langle w_B, \phi_B(x_i)\rangle - b_B| \leq \eta_i + \epsilon \\
& y_i \left(\langle w_A, \phi_A(x_i)\rangle + b_A\right) \geq 1 - \xi_i^A \\
& y_i \left(\langle w_B, \phi_B(x_i)\rangle + b_B\right) \geq 1 - \xi_i^B \\
& \xi_i^A \geq 0, \quad \xi_i^B \geq 0, \quad \eta_i \geq 0 \quad \text{all for} \quad 1 \leq i \leq \ell.
\end{aligned}$$

The final SVM-2K decision function is then $h(x) = \text{sign}(f(x))$, where

$$f(x) = 0.5\left(\left\langle \hat{w}_A, \phi_A(x)\right\rangle + \hat{b}_A + \left\langle \hat{w}_B, \phi_B(x)\right\rangle + \hat{b}_B\right) = 0.5\left(f_A(x) + f_B(x)\right). \quad (1)$$

B Supplementary results

See Tables 5 and 6.

Table 5 Results in the biased setting

	TFIDF	W_{lin}	$W_{\text{text}}(4)$	$W_{\text{eye+text}}(26)$	$W_{\text{us}}(26)$	SVM_{ex}	SVM-2K
Astronomy	53.8	13.5	33.3	15.8	20.0	58.0	57.4
Ball games	100.0	83.2	48.7	81.2	37.9	100.0	100.0
Cities	93.5	32.0	15.1	46.2	37.8	100.0	100.0
Court systems	71.6	44.0	19.0	36.0	44.3	74.8	71.3
Dinosaurs	99.1	59.7	24.0	77.9	42.6	86.0	87.4
Education	82.6	30.6	17.7	46.1	38.6	73.3	82.4
Elections	81.1	61.3	29.2	68.8	34.5	76.0	83.6
Family	44.0	6.7	12.0	16.8	5.7	65.0	66.7
Film	56.1	16.7	11.0	33.4	17.7	68.6	59.2
Government	54.8	30.8	8.9	41.5	25.4	55.5	53.4
Internet	47.0	44.8	9.1	44.6	11.7	49.7	55.0
Languages	90.6	60.2	23.8	86.7	52.6	95.0	95.0
Literature	40.3	15.5	7.2	12.1	11.5	26.8	34.2
Music	68.4	8.0	10.4	11.7	11.7	72.4	74.2
Natural disasters	81.9	62.9	29.1	66.3	34.6	99.1	98.3
Olympics	88.6	53.4	10.0	70.2	45.3	76.4	80.0
Optical devices	93.3	26.6	26.4	38.7	23.2	85.2	91.6
Postal system	90.6	50.5	11.5	76.9	66.3	94.3	95.0
Printing	84.2	63.4	42.2	69.0	63.7	86.2	85.5
Sculpture	69.5	34.9	9.8	31.6	25.7	80.1	79.5
Space exploration	88.4	49.5	6.4	67.6	42.8	88.5	89.7
Speeches	79.0	26.5	19.4	40.7	19.2	84.8	84.8
Television	63.5	26.8	33.4	32.3	28.0	71.2	73.8
Transportation	61.8	25.6	26.6	27.5	18.0	66.1	66.5
Writing systems	51.6	17.2	14.3	18.0	23.6	45.4	60.3
Average	73.4	37.8	19.9	46.3	31.3	75.1	77.0

These have been computed on a test set of 250 documents, with ten documents being positive

Table 6 Results in the unbiased setting

User	Topic	Pos/test set size	Random	TFIDF	W_{lin}	$W_{\text{text}}(4)$	$W_{\text{eye+text}}(26)$	SVM_{ex}	SVM-2K
1	Ball games	10/50	25.7	30.7	24.3	27.0	30.2	99.0	99.0
1	Dinosaurs	10/50	25.7	35.9	34.5	30.4	29.3	88.6	88.7
1	Space exploration	10/50	25.7	34.1	31.9	22.0	29.3	76.0	76.1
1	Literature	10/60	21.9	12.3	16.3	15.6	18.3	59.0	59.3
1	Government	10/50	25.7	22.4	29.9	18.5	26.4	67.5	67.8
1	Court systems	10/60	21.9	50.1	56.4	30.0	51.3	72.1	71.9
1	Cities	10/50	25.7	34.8	42.1	20.9	40.3	82.6	82.6
1	Film	10/60	21.9	23.3	16.3	18.8	16.5	77.3	77.1

Table 6 continued

User	Topic	Pos/test set size	Random	TFIDF	W_{lin}	$W_{\text{text}}(4)$	$W_{\text{eye+text}}$ (26)	SVM_{ex}	SVM-2K
1	Sculpture	10/60	21.9	14.8	23.0	15.8	16.4	77.7	78.1
1	Natural disasters	10/60	21.9	20.3	17.5	13.5	16.8	88.4	88.5
2	Education	10/40	31.3	28.0	21.6	23.5	30.8	86.1	86.2
2	Printing	10/40	31.3	46.6	34.0	33.0	34.1	87.1	86.9
2	Writing systems	10/50	25.7	17.0	15.5	15.6	15.2	85.7	85.5
2	Optical devices	10/40	31.3	23.2	25.1	32.7	23.0	89.0	89.1
2	Internet	10/60	21.9	12.8	11.8	12.0	12.4	57.2	57.2
2	Family	10/40	31.3	26.7	17.5	24.5	17.8	76.8	76.8
2	Television	10/50	25.7	17.0	18.1	23.7	17.6	77.7	77.7
2	Speeches	10/40	31.3	26.1	35.9	38.1	33.8	86.0	85.9
2	Postal system	10/60	21.9	24.1	21.7	22.0	26.7	94.7	94.7
2	Languages	10/70	19.1	35.4	31.4	51.3	28.2	99.0	99.0
3	Music	10/60	21.9	26.9	20.8	14.3	26.2	55.8	55.4
3	Olympics	10/50	25.7	39.1	54.4	30.9	43.0	93.5	93.7
3	Astronomy	10/50	25.7	19.2	20.5	24.3	18.4	77.0	77.0
3	Optical devices	10/40	31.3	23.2	21.9	31.7	23.8	89.0	89.1
3	Internet	10/60	21.9	12.8	11.5	12.0	11.3	57.2	57.2
3	Family	10/40	31.3	26.7	28.9	24.9	27.4	76.8	76.8
3	Television	10/50	25.7	17.0	21.7	23.8	23.9	77.7	77.7
3	Speeches	10/40	31.3	26.1	45.0	37.8	47.2	86.0	85.9
3	Postal system	10/60	21.9	24.1	24.2	22.2	24.4	94.7	94.7
3	Languages	10/70	19.1	35.4	44.1	53.1	30.7	99.0	99.0
4	Music	10/40	31.3	28.8	24.1	20.0	25.6	83.5	83.6
4	Olympics	10/60	21.9	30.0	56.8	19.6	53.6	89.0	89.1
4	Literature	10/40	31.3	25.4	25.2	22.5	25.7	72.7	72.7
4	Dinosaurs	10/70	19.1	14.6	23.9	23.0	19.3	81.8	81.5
4	Internet	10/40	31.3	22.2	30.0	26.2	31.6	74.1	74.1
4	Natural disasters	10/50	25.7	19.8	23.0	18.2	24.3	88.1	88.4
4	Court systems	10/40	31.3	38.9	43.5	23.7	49.1	84.0	84.1
4	Education	10/70	19.1	19.3	24.3	18.9	29.1	91.3	91.2
4	Speeches	10/60	21.9	14.6	17.9	15.9	15.2	81.1	81.2
4	Space exploration	10/40	31.3	29.0	30.0	23.9	35.6	92.6	92.7
5	Writing systems	10/40	31.3	23.2	27.0	32.7	33.2	86.0	86.1
5	Postal system	10/70	19.1	22.8	13.7	17.4	16.0	94.5	94.4
5	Transportation	10/50	25.7	24.4	28.9	21.8	27.7	67.8	67.7
5	Languages	10/30	40.2	69.7	66.0	74.1	64.7	99.9	99.9
5	Cities	10/50	25.7	19.3	21.4	20.1	18.5	84.9	84.8
5	Elections	10/30	40.2	54.0	41.2	51.7	43.5	89.3	89.3
5	Astronomy	10/40	31.3	17.3	18.1	18.9	19.2	81.3	81.1

Table 6 continued

User	Topic	Pos/test set size	Random	TFIDF	W_{lin}	$W_{\text{text}}(4)$	$W_{\text{eye+text}}$ (26)	SVM_{ex}	SVM-2K
5	Government	10/60	21.9	24.0	33.2	35.9	26.5	60.3	60.1
5	Optical devices	10/50	25.7	26.7	16.5	19.4	16.5	79.2	79.4
5	Sculpture	10/40	31.3	26.2	24.9	31.4	26.2	85.8	85.8
6	Speeches	10/40	31.3	29.7	54.3	23.3	38.6	87.5	87.5
6	Family	10/50	25.7	13.7	13.0	14.8	12.9	78.3	78.3
6	Education	10/50	25.7	44.7	36.7	27.0	39.9	88.9	89.0
6	Printing	10/60	21.9	18.1	46.9	17.8	38.3	91.4	91.0
6	Space exploration	10/70	19.1	33.8	30.6	18.0	24.9	67.6	67.6
6	Writing systems	10/60	21.9	13.4	23.8	17.3	21.4	52.3	52.1
6	Music	10/80	16.9	11.8	16.9	11.6	17.5	71.6	72.0
7	Transportation	10/50	25.7	18.6	20.5	20.5	20.4	73.1	73.1
7	Ball games	10/50	25.7	43.8	22.3	40.1	22.7	98.7	98.7
8	Optical devices	10/60	21.9	12.5	14.5	13.6	14.7	85.4	85.3
8	Astronomy	10/70	19.1	14.0	15.2	14.3	10.8	67.4	67.4
8	Television	10/40	31.3	27.0	31.9	28.5	32.5	79.9	79.9
8	Film	10/50	25.7	21.0	25.1	20.2	27.1	69.3	69.4
8	Literature	10/30	40.2	27.5	37.8	26.9	37.1	72.1	72.1
8	Dinosaurs	10/50	25.7	28.1	22.3	21.3	24.8	95.9	96.0
8	Space exploration	10/70	19.1	24.3	38.9	14.6	29.4	87.1	87.2
8	Transportation	10/40	31.3	21.6	27.6	36.7	26.5	74.6	74.5
8	Sculpture	10/70	19.1	12.2	19.0	13.6	18.5	70.5	70.5
8	Government	10/50	25.7	14.0	21.8	18.0	24.7	76.3	75.9
9	Television	10/50	25.7	16.2	19.8	21.1	21.3	65.8	66.0
9	Optical devices	10/70	19.1	11.9	15.8	12.0	18.7	56.4	56.4
9	Olympics	10/50	25.7	49.3	44.6	25.6	51.3	90.5	90.6
9	Internet	10/50	25.7	26.0	19.5	17.1	17.0	64.2	64.8
9	Speeches	10/60	21.9	14.1	16.7	13.9	14.2	89.0	89.2
9	Ball games	10/60	21.9	30.2	46.9	40.1	40.1	93.5	93.2
9	Film	10/40	31.3	31.1	43.0	21.0	37.1	79.5	79.5
9	Cities	10/60	21.9	14.9	15.0	17.0	16.0	71.5	71.5
9	Education	10/50	25.7	24.5	35.9	37.8	28.2	77.9	78.4
9	Natural disasters	10/50	25.7	25.1	21.8	20.9	21.6	70.5	70.4
10	Court systems	10/60	21.9	41.9	16.4	19.7	18.4	80.6	80.5
10	Elections	10/60	21.9	32.9	16.6	42.7	18.7	88.9	88.9
10	Printing	10/50	25.7	40.3	67.4	36.6	66.9	92.4	92.3
10	Languages	10/60	21.9	37.7	58.7	22.1	44.2	94.9	94.9
10	Writing systems	10/70	19.1	11.9	17.2	17.4	14.9	70.9	70.7
10	Literature	10/60	21.9	12.6	14.8	13.9	14.1	41.0	40.8
10	Music	10/30	40.2	48.2	32.3	47.6	34.8	81.3	81.3

Table 6 continued

User	Topic	Pos/test set size	Random	TFIDF	W_{lin}	$W_{\text{text}}(4)$	$W_{\text{eye+text}}$ (26)	SVM_{ex}	SVM-2K
10	Astronomy	10/60	21.9	16.8	16.2	24.5	15.5	44.5	44.4
10	Postal system	10/40	31.3	26.3	30.1	30.9	39.6	93.5	93.5
	Average		25.7	25.9	28.1	24.8	27.4	80.0	80.0

The size of the test set and the proportion of positive documents in it vary from session to session (third column)

References

Budzik, J. Hammond, K.J.: User interactions with everyday applications as context for just-in-time information access. In: 5th International Conference on Intelligent User Interfaces, pp. 44–51. ACM, New Orleans (2000)

Claypool, M., Le, P., Wased, M., Brown, D.: Implicit interest indicators. In: 6th International Conference on Intelligent User Interfaces, pp. 33–40. Santa Fe (2001)

Conati, C., Mertena, C.: Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. Knowl. Based Syst. **20**, 557–574 (2007)

Czerwinski, M., Dumais, S., Robertson, G., Dziadosz, S., Tiernan, S., van Dantzich, M.: Visualizing implicit queries for information management and retrieval. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 560–567. Pittsburgh (1999)

Dhanjal, C., Gunn, S.R., Shawe-Taylor, J.: Sparse feature extraction using generalised partial least squares. In: IEEE International Workshop on Machine Learning for Signal Processing, pp. 27–32. Maynooth (2006)

DiCiccio, T.J., Efron, B.: Bootstrap confidence intervals. Stat. Sci. **11**, 189–228 (1996)

Dumais, S., Cutrell, E., Sarin, R., Horvitz, E.: Implicit queries (IQ) for contextualized search. In: 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 594. ACM, Sheffield (2004)

Farquhar, J.D.R., Hardoon, D.R., Meng, H., Shawe-Taylor, J., Szedmak, S.: Two view learning: SVM-2K, theory and practice. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems, pp. 355–362. MIT Press, Cambridge (2006)

Fono, D., Vertegaal, R.: EyeWindows: evaluation of eye-controlled zooming windows for focus selection. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 151–160. ACM Press, Portland (2005)

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

Granaas, M.M., McKay, T.D., Laham, R.D., Hurt, L.D., Juola, J.F.: Reading moving text on a CRT screen. Hum. Factors **26**, 97–104 (1984)

Hardoon, D.R., Szedmak, S.R., Shawe-Taylor, J.R.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**, 2639–2664 (2004)

Hardoon, D.R., Ajanki, A., Puolamaki, K., Shawe-Taylor, J., Kaski, S.: Information retrieval by inferring implicit queries from eye retrieval by inferring implicit queries from eye movements. In: 11th Intelligence and Statistics. San Juan. Electronic proceedings at http://www.stat.umn.edu/~aistat/proceedings/start.htm (2007)

Howard, D.L., Crosby, M.E.: Snapshots from the eye: towards strategies for viewing biblographic citations. In: Savendy, G., Smith, M.J. (eds.) Human-Computer Interaction: Software and Hardware Interfaces, pp. 488–493. Elsevier, Amsterdam (1993)

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 154–161. ACM, Salvador (2005)

Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. ACM SIGIR Forum **37**, 18–28 (2003)

King, L.: The relationship between scene and eye movements. In: 35th Annual Hawaii International Conference on System Sciences, p. 136b. Big Island (2002)
Levy-Schoen, A., O'Regan, K. : The control of eye movements in reading. In: Kolers, P.A., Wrolstad, M.E., Bouma, H. (eds.) Processing of visible language, pp. 7–36. Plenum Press, New York (1979)
Maglio, P.P., Campbell, C.S.: Attentive agents. Commun. ACM **46**, 47–51 (2003)
Maglio, P.P., Barrett, R., Campbell, C.S., Selker, T.: SUITOR: an attentive information system. In: International Conference on Intelligent User Interfaces, pp. 169–176. ACM, New Orleans (2000)
Meng, H., Hardoon, D.R., Shawe-Taylor, J., Szedmak, S.: Generic object recognition by distinct features combination in machine learning. In: 17th Annual Symposium on Electronic Imaging, pp. 90–98. San Jose (2005)
Puolamäki, K., Kaski, S. (eds.): Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling. Helsinki University of Technology, Otaniemi. http://www.cis.hut.fi/inips2005/ (2006)
Puolamäki, K., Salojärvi, J., Savia, E., Simola, J. Kaski, S.: Combining eye movements and collaborative filtering for proactive information retrieval. In: 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 146–153. Salvador, (2005)
Rafter, R., Smyth, B.: Passive profiling from server logs in an online recruitment environment. In: Workshop on Intelligent Techniques for Web Personalization, pp. 35–41. Seattle (2001)
Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychol. Bull. **124**, 372–422 (1998)
Rosipal, R., Trejo, L.J.: Kernel partial least squares regression in reproducing kernel Hilbert space. J. Mach. Learn. Res. **2**, 97–123 (2001)
Salojärvi, J., Kojo, I., Simola, J., Kaski, S.: Can relevance be inferred from eye movements in information retrieval?. In: Workshop on Self-Organizing Maps. Kyushu Institute of Technology, Hibikino, pp. 261–266. Kitakyushu (2003)
Salojärvi, J., Puolamäki, K., Kaski, S.: Implicit relevance feedback from eye movements. In: Duch, W., Kacprzyk, J., Oja, E., Zadrozny, S. (eds.) Artificial Neural Networks: Biological Inspirations—ICANN 2005, Lecture Notes in Computer Science 3696, pp. 513–518. Springer, Berlin (2005a)
Salojärvi, J., Puolamäki, K., Simola, J., Kovanen, L., Kojo, I. Kaski, S.: Inferring relevance from eye movements: feature extraction. Technical Report A82, Helsinki University of Technology, Publications in Computer and Information Science. http://www.cis.hut.fi/eyechallenge2005/ (2005b)
Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. Vol. 1, McGraw-Hill, New York (1983)
Turpin, A., Scholer, F.: User performance versus precision measures for simple search tasks. In: 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 11–18. Seattle (2006)
Ward, D.J., MacKay, D.J.C.: Fast hands-free writing by gaze direction. Nature **418**, 838 (2002)

Author Biographies

Antti Ajanki is a Ph.D. candidate in Computer Science at Helsinki University of Technology. He received his M.Sc. (Tech.) degree in Computer Science from Helsinki University of Technology in 2006. His research interests include machine learning and information retrieval. Previous research has included work in the field of bioinformatics.

Dr. David R. Hardoon is a research fellow at the University College, London. He is currently working on projects that are focused on learning the structure of music, medical analysis, multilingual and multi-modal integration. He has a keen interest in multi-view learning, kernel methods, regression, and sparsity. He has previously worked on various research projects in the fields of Taxonomy, Image analysis, classification and content based retrieval systems. David received his first class B.Sc. Hons. in Computer Science with Artificial Intelligence from the Royal Holloway, University of London and his PhD in Computer Science in the field of Machine Learning from the University of Southampton. He has also received the PhD PASCAL label award from his active participation in the PASCAL network (More information about him can be found at http://www.DavidRoiHardoon.com/).

Prof. Samuel Kaski is Professor of Computer Science at Helsinki University of Technology, vice director of the Adaptive Informatics Research Centre, a center of excellence of the Academy of Finland, and group leader in Helsinki Institute for Information Technology HIIT. Dr. Kaski received his M.Sc. and D.Sc. (PhD) degrees in Computer Science from Helsinki University of Technology. His main research field is statistical machine learning, with applications in bioinformatics and information retrieval. He has authored 130 refereed papers in these fields.

Dr. Kai Puolamäki is a lecturing researcher (assistant professor) in the Department of Computer Science at the Helsinki University of Technology. He completed his Ph.D. in 2001 in theoretical physics from the University of Helsinki. His primary interests lie in the areas of data mining, machine learning and related algorithms, and especially their application in ecology and user modelling.

Prof. John Shawe-Taylor has been the Director of the Centre for Computational Statistics and Machine Learning at University College, London since July 2006. He obtained a PhD in Mathematics at Royal Holloway, University of London in 1986. He subsequently completed an MSc in the Foundations of Advanced Information Technology at Imperial College. He was promoted to Professor of Computing Science in 1996. He has published over 150 research papers. He led the ISIS research group at the University of Southampton from 2003 to 2006. He is the scientific coordinator of the EC funded Network of Excellence PASCAL.