Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola, and Samuel Kaski. 2005. Combining eye movements and collaborative filtering for proactive information retrieval. In: Gary Marchionini, Alistair Moffat, John Tait, Ricardo Baeza-Yates, and Nivio Ziviani (editors). Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005). Salvador, Brazil. 15-19 August 2005. New York, USA. ACM Press, pages 146-153.

http://doi.acm.org/10.1145/1076034.1076062

Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval

Kai Puolamäki
[1]Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400
FIN-02015 HUT, Finland
Kai.Puolamaki@hut.fi

Jarkko Salojärvi
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400
FIN-02015 HUT, Finland
Jarkko.Salojarvi@hut.fi

Eerika Savia
Laboratory of Computer and Information Science
Helsinki University of Technology
P.O. Box 5400
FIN-02015 HUT, Finland
Eerika.Savia@hut.fi

Jaana Simola
Center for Knowledge and Innovation Research
Helsinki School of Economics and Business Administration
Tammasaarenkatu 3
FIN-00180 Helsinki, Finland
Jaana.Simola@hkkk.fi

Samuel Kaski[1,2]
[2]Department of Computer Science
P.O. Box 68
FIN-00014 University of Helsinki, Finland
Samuel.Kaski@hut.fi

ABSTRACT

We study a new task, proactive information retrieval by combining implicit relevance feedback and collaborative filtering. We have constructed a controlled experimental setting, a prototype application, in which the users try to find interesting scientific articles by browsing their titles. Implicit feedback is inferred from eye movement signals, with discriminative hidden Markov models estimated from existing data in which explicit relevance feedback is available. Collaborative filtering is carried out using the User Rating Profile model, a state-of-the-art probabilistic latent variable model, computed using Markov Chain Monte Carlo techniques. For new document titles the prediction accuracy with eye movements, collaborative filtering, and their combination was significantly better than by chance. The best prediction accuracy still leaves room for improvement but shows that proactive information retrieval and combination of many sources of relevance feedback is feasible.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*; I.5.1 [**Computing Methodologies**]: Pattern Recognition—*Models*; H.1.2 [**Information Systems**]: Models and Principles—*User/Machine Systems*

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

General Terms

Algorithms, Experimentation, Theory

Keywords

Collaborative filtering, eye movements, hidden Markov model, latent variable model, mixture model, proactive information retrieval, relevance feedback

1. INTRODUCTION

In a typical information retrieval setup users formulate queries that express their interests. The task of an information retrieval system then is to identify documents that best match the query terms, based on the contents of the documents. Alternatively, documents can be used as very complex queries to find other relevant documents, when similarity measures have been defined between documents. The systems may additionally collect explicit relevance feedback from the user, by asking which of the retrieved documents were relevant, and combine the results to a new search.

Information retrieval systems would be much more user-friendly if the number of successive explicit queries and explicit relevance evaluations could be reduced—or eliminated altogether. Our work is a feasibility study on how far we can go by measuring implicit feedback signals from the user, and by combining them with existing data about preferences of a group of similar-minded users. The task is to predict relevance; if the predictions are successful they can be used in a variety of proactive applications, including proactive information retrieval.

We infer user interest from eye movements with probabilistic models that predict whether a user finds a text relevant, given her eye movement trajectory while reading the text. The key assumption motivating the use of eye movements is that attention patterns correlate with relevance,

and that attention patterns are reflected in eye movements (see [19]). At the simplest, people tend to pay more attention to objects they find relevant or interesting.

Gaze direction is an indicator of the focus of attention, since accurate viewing is possible only in the central fovea area (only 1–2 degrees of visual angle) where the density of photoreceptor cells is highly concentrated. A detailed inspection of a scene is carried out in a sequence of *saccades* (rapid eye movements) and *fixations* (during which the eye is fairly motionless). Information about the environment is mostly gathered during fixations. The physiology suggests that eye movements can provide a rich source of information about the attention and interest patterns of the user. Indeed, psychologists have studied eye movements as an indicator of different cognitive processes for decades [18], and a recent feasibility study [19] showed that relevance can be inferred from eye movements, at least to a certain degree.

Our key contribution is that we do not assume anything about the details of this relationship between the attention and eye movement patterns; we infer everything we need from data, using machine learning methods.

Collaborative filtering is another, complementary source of relevance information. The goal of collaborative filtering is to predict the relevance of a document to a given user, based on a database of explicit or implicit relevance ratings from a large population of users. In this work we complement the rich but noisy eye movement-based relevance feedback with collaborative filtering, using a probabilistic latent variable model.

Finally, we combine the predictions from eye movements and collaborative filtering, again with a probabilistic model. The system is modular in the sense that new better components can easily be plugged in later, to replace the ones we use in this feasibility study.

In our prototype application the users browse titles of scientific articles and their eye movements are measured. We then combine the relevance predictions of the collaborative filtering model with the model that predicts the relevance from implicit feedback information.

The main research questions of this paper are:

1. How does the eye movement model perform in inferring which articles are interesting?
2. How does the collaborative filtering model perform in the same task?
3. How do the models compare against each other?
4. Is it feasible to combine relevance predictions from implicit feedback (eye movements), and other sources (collaborative filtering), and how to do the combination?

2. RELATED WORK

To our knowledge, the combination of using implicit feedback from eye movements and relevance prediction from a collaborative filtering model is new. However, earlier work exists in several separate fields: inferring relevance implicitly from eye movements [19], extending queries or modeling user preferences by estimating relevance from implicit feedback [7], and using user modeling to determine documents that may be relevant to a group of users [4, 12]. There have also been various studies on combining collaborative filtering and content-based filtering in general (e.g. [1, 14]).

Eye movements have earlier been utilized as alternative input modalities for either pointing at icons or typing text in human-computer interfaces (the most recent application being [25]). The first application where user interest was inferred from eye movements was an interactive story teller [23]. The story teller concentrated more on items that the user was gazing at on a display. Rudimentary relevance determination is needed also in [5], where a proactive translator is activated if the reader encounters a word which she has difficulties in understanding. These difficulties are inferred from eye movements.

Traditionally implicit feedback in IR has been derived from document reading time, or by monitoring user behavior: saving, printing, and selecting of documents (see [7] for a good overview on different approaches). Use of eye movements as a source of implicit feedback for IR is a relatively new concept. A prototype attentive agent application (Simple User Interest Tracker, Suitor) is introduced in [10, 11]. The agent 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, the document is defined relevant, and more information on the topic is sought and displayed. The feasibility of the application was not experimentally verified, however. To our knowledge the only study with statistically tested significance is [19, 20], which is a simple feasibility study. The experimental setup is close to ours, but the task is much simpler. The user is presented a question and a list of possible answers, some of which are relevant to the question and one provides the answer. The eye movements are then used to infer the relevant lines, as well as the correct answer.

Traditionally collaborative filtering has been performed by memory-based techniques, in which one first identifies users similar to a given user and then gives predictions based on interests of those users (see e.g. GroupLens [8], or Ringo [22]). However, the time and memory requirements of the memory-based techniques do not scale well as the number of users and documents increases.

We propose a model-based approach, which is based on the User Rating Profile model (URP) [12]. We have extended the previous work by optimizing the URP by using Markov Chain Monte Carlo (MCMC) integration instead of the variational approximation used earlier. The model structure of URP is closely related to a probabilistic latent variable model introduced by Pritchard et al. [16], and Latent Dirichlet Allocation (LDA) [2] which is also known as Multinomial PCA (mPCA) [3].

3. MODELS

3.1 Eye Movement Modeling

Eye movements were modeled using hidden Markov models (HMMs); they are simple yet reliable models for sequential data. We use two kinds of HMMs: ordinary and discriminative, both modeling word-level eye movement data. In eye movement research, HMMs have earlier been used for segmenting the low-level eye movement signal to detect focus of attention and for implementing (fixed) models of cognitive processing [21]. Discriminative HMMs have been previously applied to eye movement data in [20].

Prediction of known classes with machine learning methods is based on a labeled data set from which the predictive model is learned. We collected such a set by measuring eye

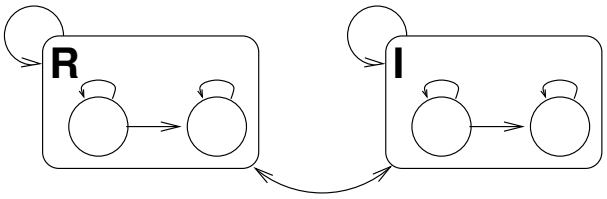

Figure 1: The topology of the discriminative hidden Markov model. The first level models transitions between sentences having relevance $r \in \{I, R\}$ and the second level (within the boxes) models transitions between the words in a sentence.

movements in a setting where relevance was known: explicit feedback for presented sentences (titles of scientific documents) was collected from the user during the recording session. The experimental setup is described in more detail in Section 4.1.

Hidden Markov Models

The simplest model that takes the sequential nature of eye movement data into account is a two-state HMM. We optimized one model individually for each of the two classes, in our case relevant (R) and irrelevant (I) sentences. In a prediction task a maximum a posteriori (MAP) estimate was computed. The two HMMs were fitted to data by the Baum-Welch (BW) algorithm that maximizes the log-likelihood of the data Y given the model and its parameters ψ, that is, $\log p(Y|\psi)$ [17]. The model is described in more detail in [20].

Discriminative Hidden Markov Models

In discriminative modeling we want to predict the relevance $r = \{I, R\}$ of a sentence, given the observed eye movements Y. Formally, we optimize

$$\log p(r|Y, \psi) \quad .$$

In speech recognition, where HMMs have been extensively used for decades, the current state-of-the-art HMMs are discriminative. The parameters of the discriminative HMM can be optimized with an Extended Baum-Welch (EBW) algorithm [15], which is a modification of the original BW.

Eye movements are modeled with a two-level discriminative HMM, where the first level models transitions between sentences whereas the second level models transitions between words within a sentence. The topology of the model is shown in Figure 1.

In our implementation, the first level Markov model has two states, each modeling one sentence class (I or R). Each state has the following exponential family emission distributions: (1) A multinomial distribution emitting the relevance of the line, r. This distribution is fixed; for each state one of the probabilities is one and the other is zero. (2) A *Viterbi distribution* emitting the probability of the sequence of words in a sentence. The Viterbi distribution is defined by the probability of a Viterbi path [24] trough a two-state Markov model forming the second level in our model. The two states of the second level model emit the exponential observation distributions. The modeled eye movement features are described in Section 4.1.

When optimizing the model the most likely path through the second level model is sought by the Viterbi approximation [24]. The discriminative Extended Baum-Welch algorithm optimizes the full model, keeping the Viterbi path in the second level model fixed.

3.2 Collaborative Filtering Model

Collaborative filtering was carried out with a state-of-the-art latent topic model, the User Rating Profile model (URP) [12]. This model was used because in [12] it outperformed several other latent topic models. It was originally optimized with variational Bayesian methods (*variational URP*). We also implemented a potentially more accurate Markov Chain Monte Carlo (MCMC) integration method to compute the predictions from the model, using Gibbs sampling (*Gibbs URP*).

Table 1: Notation

Symbol	Description
u	user index
d	document index
r	binary relevance
Z	user group index
N_U	number of users
N_D	number of documents
K_U	number of user groups
P_{urp}	relevance prediction of the URP
P_{eye}	relevance prediction of the HMM

User Rating Profile Model

URP is a generative model which generates a binary rating r for a given (user, document) pair.[1] Our notation is summarized in Table 1. We estimate the posterior distribution $P(r|u, d, \mathcal{D})$ by Gibbs sampling where $\mathcal{D}$ denotes the training data that consists of observations (u, d, r). The model assumes that there are a number of latent user groups whose preferences on the documents vary, and the users belong to these groups probabilistically. Alternatively, the groups can be interpreted as different "attitudes" of the user, and the attitude may be different for different documents.

The generative process proceeds according to the following steps (see also Figure 2):

- For each user, a vector of multinomial parameters $\boldsymbol{\theta}(u)$ is drawn from Dirichlet($\boldsymbol{\alpha}$).[2] The parameter vector $\boldsymbol{\theta}(u)$ contains the probabilities for the user to have different attitudes Z, i.e., to belong to different user groups Z.
- For each user u, a user group or attitude Z is drawn for each document d, from the user's Multinomial($\boldsymbol{\theta}(u)$). The value of Z in effect selects the parameters $\boldsymbol{\beta}(Z, d)$ from the set of parameters in the node labeled by β in Figure 2.
- For each (user group, document) pair (Z, d), a vector of binomial parameters $\beta(Z, d)$ is drawn from Dirichlet($\boldsymbol{\alpha}_\beta(Z, d)$). The parameters $\beta(Z, d)$ define the probability of the user group Z to consider document d relevant (or irrelevant).

[1]Note that the model allows also multiple-valued ratings if the binomial is replaced with a multinomial.
[2]We denote distributions with capitalized words followed by their parameters in parentheses.

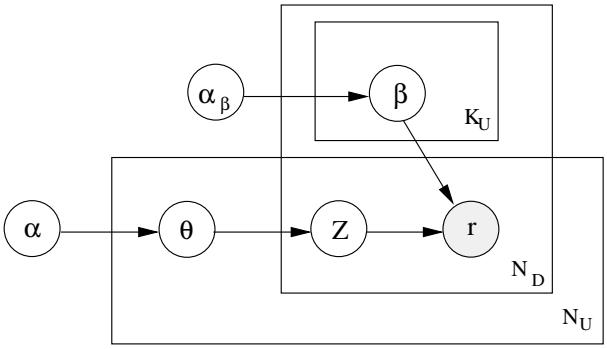

Figure 2: A graphical model representation of URP. The grey circle indicates an observed value. The boxes are "plates" representing replicates and the index at the bottom right corner of each plate indicates the number of replicates. The lowest plate, labeled with N_U, represents users. The plate labeled with N_D represents the repeated choice of user group and document. The plate labeled with K_U represents the multinomial models of relevance for the different user groups.

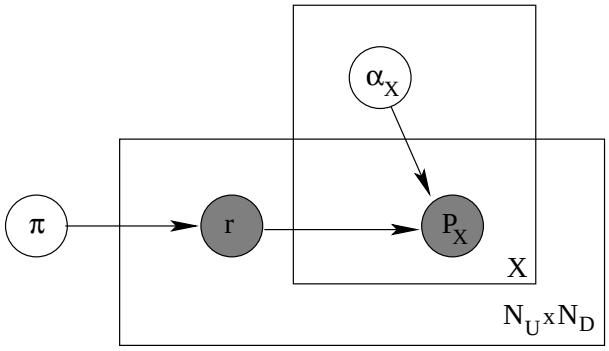

Figure 3: A graphical model representation of the discriminative Dirichlet mixture model. X is the index of the model that predicts relevance, in our case $X \in \{eye, urp\}$. The grey circles indicate observed values. In our model we observe triplets (r, P_{eye}, P_{urp}) for each user-document pair.

- For each pair (Z, d), a binary relevance value r is drawn from the Binomial$(\beta(Z, d))$.

Comparison to Other Latent Topic Models

In the URP model each user is assigned a distribution of multinomial parameters $\boldsymbol{\theta}$ and the latent user group ("topic" in text modeling) Z is sampled repeatedly for each document. A user can therefore belong to many groups with varying degrees. In URP, the multinomial parameters $\boldsymbol{\theta}$ are marginalized out from the maximum likelihood cost function. In the well-known latent topic model called Probabilistic Latent Semantic Analysis [4], the number of parameters grows with the number of users, since each user is given a fixed set of multinomial parameters $\boldsymbol{\theta}$.

URP is closely related to Pritchard's latent variable model [16] and Latent Dirichlet Allocation (LDA) [2] (also known as multinomial PCA). URP can be seen as an extension to LDA with one extra dimension in the parameter matrix $\boldsymbol{\beta}$ to represent the possible different rating values. In our case we only have two values.

Evaluating Gibbs URP

In Gibbs URP a five-fold cross-validation within the training set was first carried out to determine the optimal number of user groups in the range $\{1, 2, \ldots, N_U\}$. In our experiments the optimal number of user groups was found to be two, which was later used when computing the predictions for the final test set.

The duration of the burn-in period was determined by running three MCMC chains in parallel and monitoring the convergence of predictions.

Dumb Model and Document Frequency Model

We introduced two simple models to give baseline results. The *dumb model* classifies all documents to the largest class, $P(r = 0) = 1$. The *document frequency model* does not take into account differences between users or user groups. It simply models the probability of a document being relevant as the frequency of $r = 1$ in the training data for the document,

$$P(r = 1 \mid d) = \frac{\sum_u \#(u, d, r = 1)}{\sum_{u,r} \#(u, d, r)} \quad .$$

3.3 Combining Models

We started by examining the prediction performance of each of the models separately. Since the models use different sources of information, the natural extension is to combine their predictions.

Both models produce a probability of relevance for each given (user, document) pair. The simplest way to combine the models is to train the models independently and combine the predicted probabilities to produce the final prediction. This approach has the advantage of being modular and easily extensible.

Discriminative Dirichlet Mixture Model

We formulated a generative model for combining probabilities. Let us denote the prediction of the collaborative filtering model by P_{urp} and the prediction of the eye movement model by P_{eye}.

We first define a model that generates the observed relevances $r \in \{0, 1\}$ and the (noisy) predictions P_{urp} and P_{eye}. Our goal is to find an expression for $P(r|P_{urp}, P_{eye}, \varphi)$, where φ denotes all parameters of the model.

The generative process of the *discriminative dirichlet mixture model* is (see Figure 3) as follows:

- For each (user, document) pair, a binary relevance r is drawn from Binomial(π).
- For each $X \in \{urp, eye\}$, a vector of multinomial (in this case binomial) parameters P_X is drawn from Dirichlet(α_X^r).

The observed variables of our model are the binary relevances r_n and the prediction probabilities $P_{X,u,d}$, where the indices u, d denote all (user, document) pairs. The parameters of the model are given by $\varphi = \{\pi, \alpha_{urp}^r, \alpha_{eye}^r\}$. We have ignored the priors of the parameters, since we assume

the prior to be flat, i.e., $P(r, P_X|\varphi) = P(r, P_X, \varphi)$, up to a normalization factor.

We optimize the model by maximum likelihood, and since the task is to predict relevances we build a discriminative model by maximizing the conditional log-likelihood of the relevances,

$$\mathcal{L} = \sum_{u,d} \log P(r_{u,d}|P_{urp,u,d}, P_{eye,u,d}, \varphi) \quad . \qquad (1)$$

Values of the parameters φ can be found using standard optimization methods, for instance gradient ascent.

Besides giving predictions of relevance, the Dirichlet mixture reveals how useful the different sources of relevance information are relative to each other. Some of the feedback channels may produce smaller prediction probabilities P_X than others for the observed relevances r. Some of the relevance feedback channels may additionally be noisy, that is, the prediction probabilities P_X for a given relevance r have a large variance. After optimization, the mixture parameters α_X^r will contain information about magnitude and noisiness of the probability predictions. The magnitude of the prediction is contained in the relative magnitudes of the Dirichlet components. The information of the noisiness is contained in the sum of the Dirichlet parameters: if the sum of Dirichlet parameters is large, $\sum_{i\in\{0,1\}} \alpha_{Xi}^r \gg 1$, the prediction probabilities P_X have smaller variance, and vice versa.

Linear Mixture Model

To serve as a baseline, we constructed a *linear mixture model* in which the final probability is a linear combination of the predictions of the various models (here of two models),

$$P(r|P_{urp}, P_{eye}, q) = qP_{urp}(r) + (1-q)\, P_{eye}(r) \quad , \quad q \in [0,1] \quad .$$

The parameter q is optimized by maximizing the conditional log-likelihood (Equation (1)) using standard optimization methods.

4. EXPERIMENTS

4.1 Experimental Setup

The test subjects were shown 80 pages, each containing titles of scientific articles. On each page the subject was instructed to choose the two most interesting titles in the order of preference.

The subjects participating in the experiment were researchers in vision research, artificial intelligence, and machine learning. The stimuli consisted of titles of scientific articles published during autumn 2004 in major journals in the fields of vision research (VR), artificial intelligence (AI), machine learning (ML), and general science (see Appendix A). On each page there was a randomly selected list of titles always containing two VR titles and one title from a general science journal. Half of the lists contained additionally one AI title and two ML titles, and half vice versa. Each list consisted of six titles, resulting in a total of 480 titles. The lists were shown in a randomized order to each subject, but the pages themselves were identical to all subjects.

Data was gathered in two different modalities. 22 of the subjects were asked to give their feedback explicitly via a web form, and three of the subjects participated in an eye movement experiment (Figure 4). In this paper we refer to these subjects as the *web-subjects* and *eye-subjects*, respectively. The web-subjects were given the full publication information of the most interesting paper as a reward to encourage them to find the truly most interesting titles.

Three of the subjects (the eye-subjects) were shown the same stimuli in a controlled setting where eye movements were recorded. In the experiment, the subject was instructed in a similar manner to choose the two most interesting titles from the list of six titles (the eye movements were measured during this part), then press "enter" to proceed to another display, and finally to type in the numbers corresponding to the interesting titles. Hence both explicit ratings and eye movement trajectories were available for these subjects. Eye movements were measured with a Tobii 1750 eye tracker with a screen resolution of 1280x1024.

4.2 Data

Randomly chosen 21 of the lists (26 %) and the corresponding ratings from the eye-subjects formed a common test data set for all the models. Seven of the titles in the test set were not read by the eye-subjects. They were discarded, leaving us a total of 371 titles in the test set. Test data set was not touched before computing the final results.

Eye movement models were trained with the remaining feedback data from the three eye-subjects.[3] For training the URP model we used the eye-subjects' explicit ratings that were not included in the test data set, and all the explicit feedback data from the web-subjects.

Eye Movement Data

For the eye-subjects, nine of the measured lists had to be discarded from the data sets for technical reasons, thus leaving a set of 71 lists where both explicit and implicit feedback was available. The explicit ratings were, however, not discarded.

The raw eye movement data (consisting of x and y coordinates of the gaze direction, measured with a sampling rate of 50 Hz) was segmented into a sequence of fixations and saccades by a window-based algorithm (software from Tobii), with a 20-pixel window size and a minimum duration of 80 ms, used for defining fixations. An example of an eye movement trajectory in a case where relevance can easily be determined is shown in Figure 5.

Feature extraction from the fixation-level data was then carried out as in [20]. Each fixation was first assigned to the nearest word, which segmented the eye movement trajectory into words. The following features were then computed from the segmented data to be modeled with hidden Markov models: (1) One or many fixations within the word (modeled with a binomial distribution). (2) Logarithm of the total fixation duration on the word (assumed to be Gaussian). (3) Reading behavior (multinomial): skip next word, go back to already read words, read next word, jump to an unread line, or the last fixation in an assignment.

Explicit Feedback Data

In the explicit feedback data all the selected titles were assumed to be relevant ($r = 1$) for the user, resulting in one third of all the ratings being "relevant." In other words, we did not model the users' preference order for the titles.

[3]The number of eye-subjects was chosen to be three for practical reasons. Three subjects were sufficient to train the HMM, to form the test set, and to obtain a statistically significant result.

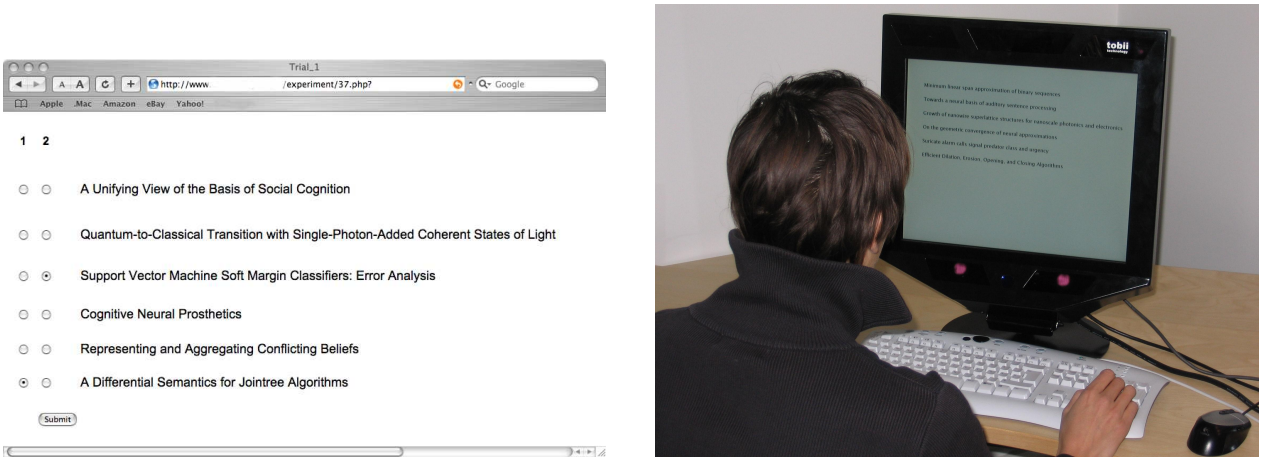

Figure 4: A total of 25 test subjects were shown 80 lists, each with six article titles. On each page the subjects chose the two most interesting titles. 22 of the subjects were asked to give their feedback explicitly via web forms (sample shown on the left). Eye movements of three subjects were measured with a Tobii 1750 eye tracker (right).

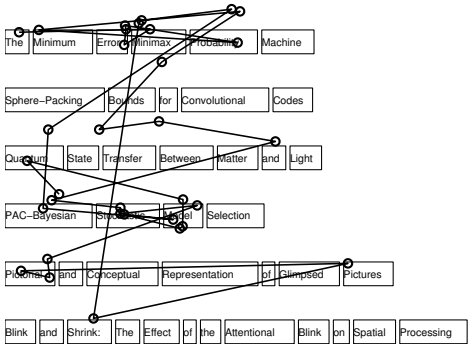

Figure 5: A reconstruction of the eye movement trajectory of a test subject during one of the assignments. Fixations are denoted by circles. The relevant sentences (R) are on lines 1 and 4. The trajectories naturally varied across the users and experiments.

5. RESULTS

For all the models we used perplexity and prediction accuracy in the test data set as measures of performance. Perplexity measures the probabilistic quality of the prediction,[4]

$$\text{perplexity} = e^{-\frac{\mathcal{L}}{N}} \ , \ \text{where } \mathcal{L} = \sum_{u,d} \log P(r_{u,d} \mid \psi) \ .$$

Here ψ denotes the parameters of the model under evaluation, the sum is taken over the test set, and N is the size of the test set. We further computed the accuracy, that is, the fraction of the items in the test data set for which the prediction was correct for all the models, and the precision and recall measures. Precision is defined as the fraction of relevance predictions that were correct. Recall is defined as the fraction of relevant items correctly predicted. The results are shown in Table 2.

The discriminative HMM produced a reasonable, though rather noisy, prediction of the relevance[5]. The difference in classification accuracy versus the dumb model was statistically significant (McNemar's test, $P \ll 0.01$). The performances of the Document Frequency Model and URP cannot be directly compared to the HMM, since the HMM prediction is only based on the eye movements from the three eye-subjects, whereas the other models utilize the explicit feedback given by the 22 web-subjects. Consequently, it is not surprising that the URP outperforms the pure HMM in terms of perplexity and accuracy measures.

As expected, URP was able to distinguish two different user groups and provide a reasonable 83 % accuracy. The accuracies of the variational version and the Gibbs version were practically equal. However, the perplexity of the Gibbs URP is better. The reason is that the variational URP finds a maximum likelihood point estimate of the model parameters, whereas the Gibbs URP integrates properly over all model parameters, resulting in a more robust probability prediction. The difference of Gibbs URP to Document Frequency Model and HMM was tested by the Wilcoxon signed rank test, applied to the negative log-likelihoods given by the models for individual test samples. The differences were significant ($P \ll 0.01$).

The discriminative Dirichlet mixture model did succeed in combining the predictions of the different models. The difference of Dirichlet Mixture Model (HMM+Gibbs URP) to Gibbs URP was statistically significant ($P \ll 0.01$) using the Wilcoxon signed rank test. The linear mixture of predictions performed poorly, placing all the weight on the prediction of the URP and ignoring the more noisy HMM al-

[4]The best possible performance yields perplexity = 1 and random guessing (coin flipping) yields perplexity = 2. If perplexity is greater than 2 the model is doing worse than random guessing. Theoretically, it can grow without a limit if the model predicts zero probability for some item in the test data set. However, we actully clipped the probabilities to the range $[e^{-10}, 1]$ implying maximum perplexity of $e^{10} \approx 22,000$

[5]The performance of simple two-state HMMs optimized for each class was similar to discriminative HMM.

Table 2: Results. Small perplexity and large accuracy, precision, and recall are better. The differences of Dirichlet Mixture Model (HMM+Gibbs URP) to Gibbs URP, as well as Gibbs URP to Document Frequency Model and HMM were tested statistically and found significant (Wilcoxon signed rank test).

Model	Perplexity	Accuracy (%)	Precision (%)	Recall (%)
Dumb Model	–	66.6	0.0	0.0
Document Frequency Model	**1.80**	69.1	55.8	35.0
Linear Mixture (HMM+Gibbs URP)	**1.50**	83.0	76.8	69.9
HMM (eye movements)	**1.78**	73.3	70.0	34.1
Variational URP (collaborative filtering)	**1.62**	83.3	77.5	69.9
Gibbs URP (collaborative filtering)	**1.50**	83.0	76.8	69.9
Dirichlet Mixture (HMM+Var. URP)	**1.56**	85.7	83.7	70.7
Dirichlet Mixture (HMM+Gibbs URP)	**1.48**	85.2	81.5	71.5

together. The precision of the best mixture (83.7 %) is quite good, taking into account that the content of documents was not modeled in any way.

Finally, we wish to point out that, for relatively small data sets, the classification accuracy is a noisy measure, as compared to perplexity. However, the difference between the accuracies of the Mixing Model (HMM+variational URP) and Gibbs URP is significant even for this noisy measure, at a moderate P-value of 0.04 (with McNemar's test). The difference between the accuracies of variational URP and the combination is not statistically significant ($P = 0.06$).

6. CONCLUSIONS AND DISCUSSION

We have set up a controlled experimental framework where the test subjects rated the relevance of titles of scientific articles. Eye movements were measured from a subset of the test subjects. The experimental setup was designed to resemble closely a real-world information retrieval scenario, where the user browses the output of, e.g., a web search engine in an attempt to find interesting documents. In our scenario a database of user preferences is combined with the measured implicit relevance feedback, resulting in more accurate relevance predictions. Collaborative filtering and implicit feedback can be used alone, or to complement standard textual content-based filtering.

We applied a discriminative time series model that produced a reasonable, though rather noisy, prediction of document relevance based on eye movement measurements. We also applied a probabilistic collaborative filtering model that produced a quite robust document relevance prediction. Thirdly, we introduced a probabilistic mixture model that can be used to combine the predictions. The mixture model clearly outperformed a simple linear method and was found necessary for making use of several information sources, the quality of which varied.

Our work provides the next step towards proactive information retrieval systems. The obvious extension is to incorporate the textual content of the documents to the models; in this work we do not utilize it at all. The second extension is to supplement or replace the eye movements by other sources of implicit feedback, such as measurements by biopotential sensors, e.g., from autonomic nervous system signals [9] or electromyographic activity [13], and respective probabilistic models.

The models for inferring relevance could also be developed further. A good opportunity is provided by our Pascal EU Network of Excellence challenge (Inferring Relevance from Eye Movements, [6]), a competition where participants are invited to develop methods that best predict relevance from eye movement data.

7. ACKNOWLEDGMENTS

The authors would like to thank the people at Tampere Unit for Computer-Human Interaction, mainly Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä for useful discussions and for providing us with the measurement time for the eye movement experiments. We would also like to thank all the persons that (more or less) voluntarily took part in our experiment.

This work was supported by the Academy of Finland, decision no. 79017, and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. The authors acknowledge that access rights to the data sets and other materials produced in the PRIMA project are restricted due to other commitments.

8. REFERENCES

[1] J. Basilico and T. Hofmann. Unifying collaborative and content-based filtering. In *Proceedings of ICML'04, Twenty-first International Conference on Machine Learning*. ACM Press, New York, 2004.

[2] D. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] W. Buntine. Variational extensions to EM and multinomial PCA. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of ECML'02, 13th European Conference on Machine Learning*, pages 23–34. Springer, Berlin, 2002.

[4] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22:89–115, 2004.

[5] A. Hyrskykari, P. Majaranta, and K.-J. Räihä. Proactive response to eye movements. In G. W. M. Rauterberg, M. Menozzi, and J. Wesson, editors, *INTERACT'03*, pages 129–136. IOS Press, 2003.

[6] `http://www.pascal-network.org/Challenges/IREM/`.

[7] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[8] J. Konstan, B. Miller, D. Maltz, and J. Herlocker. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[9] C. L. Lisetti and F. Nasoz. Maui: a multimodal affective user interface. In *MULTIMEDIA'02: Proceedings of the tenth ACM international conference on Multimedia*, pages 161–170. ACM Press, 2002.

[10] P. P. Maglio, R. Barrett, C. S. Campbell, and T. Selker. Suitor: an attentive information system. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, pages 169–176. ACM Press, 2000.

[11] P. P. Maglio and C. S. Campbell. Attentive agents. *Communications of the ACM*, 46(3):47–51, 2003.

[12] B. Marlin. Modeling user rating profiles for collaborative filtering. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[13] T. Partala, V. Surakka, and T. Vanhala. Person-independent estimation of emotional experiences from facial expressions. In *IUI'05: Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 246–248. ACM Press, 2005.

[14] A. Popescul, L. Ungar, D. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of UAI-2001*, pages 437–444. Morgan Kaufmann, 2001.

[15] D. Povey, P. Woodland, and M. Gales. Discriminative map for acoustic model adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, volume 1, pages 312–315, 2003.

[16] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–59, 2000.

[17] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[18] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422, 1998.

[19] J. Salojärvi, I. Kojo, J. Simola, and S. Kaski. Can relevance be inferred from eye movements in information retrieval? In *Proceedings of WSOM'03, Workshop on Self-Organizing Maps*, pages 261–266. Kyushu Institute of Technology, Kitakyushu, Japan, 2003.

[20] J. Salojärvi, K. Puolamäki, and S. Kaski. Relevance feedback from eye movements for proactive information retrieval. In J. Heikkilä, M. Pietikäinen, and O. Silvén, editors, *Workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004)*, Oulu, Finland, 2004.

[21] D. D. Salvucci and J. R. Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16:39–86, 2001.

[22] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating 'word of mouth'. In *Proceedings of Computer Human Interaction*, pages 210–217, 1995.

[23] I. Starker and R. A. Bolt. A gaze-responsive self-disclosing display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3–10. ACM Press, 1990.

[24] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

[25] D. J. Ward and D. J. MacKay. Fast hands-free writing by gaze direction. *Nature*, 418:838, 2002.

APPENDIX

A. LIST OF JOURNALS

Pool 1: Machine learning journals:

- Journal of Machine Learning Research
- Machine Learning
- Journal of the Royal Statistical Society: Series B (Statistical Methodology)
- IEEE Transactions on Neural Networks

Pool 2: Vision research journals:

- Vision Research
- Perception
- Journal of Experimental Psychology, Human Perception and Performance
- Trends in Cognitive Sciences

Pool 3: Artificial intelligence journals:

- Journal of Artificial Intelligence Research
- Artificial Intelligence
- IEEE Transactions on Information Theory
- IEEE Transactions on Pattern Analysis and Machine Intelligence

Pool 4: General scientific journals:

- Letters to Nature
- Science