Learning to Learn Implicit Queries from Gaze Patterns

Kai Puolamäki KAI.PUOLAMAKI@TKK.FI
Antti Ajanki ANTTI.AJANKI@TKK.FI
Samuel Kaski SAMUEL.KASKI@TKK.FI
Helsinki Institute for Information Technology, Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland

Abstract

In the absence of explicit queries, an alternative is to try to infer users' interests from implicit feedback signals, such as clickstreams or eye tracking. The interests, formulated as an implicit query, can then be used in further searches. We formulate this task as a probabilistic model, which can be interpreted as a kind of transfer or meta-learning. The probabilistic model is demonstrated to outperform an earlier kernel-based method in a small-scale information retrieval task.

1. Introduction

The classic problem in information retrieval (IR) is to rank a set of documents according to the user's current interest, with the documents most relevant for the user ranked among the first. The same theme recurs currently in other applications of machine learning such as recommender systems. Current IR systems rely mostly on explicit, typed queries to perform the ranking.

The main problem in this traditional IR scenario is that it is difficult even for experienced users to formulate good textual queries (Turpin & Scholer, 2006), and therefore user's interest needs to be inferred partly from other sources. A straightforward way is to collect explicit feedback, that is, the user labels some of the documents relevant or irrelevant for her interests. Giving explicit feedback is however laborious. It would be ideal if the IR system would be able to unobtrusively collect and use *implicit feedback* to infer the interest of the user while she works and use this information to improve the quality of the search results. We call this task *proactive information retrieval*.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

Several forms of implicit feedback, such as clickstream data, time spent during reading, and amount of scrolling and exit behaviour, have been used with some success (Kelly & Teevan, 2003; Claypool et al., 2001; Fox et al., 2005; Joachims et al., 2005; Joachims & Radlinski, 2007). While these sources of feedback are often readily available, they offer only limited information of users' interests.

Gaze patterns are a promising source of information about the attention of the user, and hence of implicit feedback. They have been used for information retrieval in two papers (Puolamäki et al., 2005; Hardoon et al., 2007). In the latter, eye tracking-based feedback improved information retrieval performance in an experiment where no explicit queries were available, and everything was inferred from the eye movements and the texts. The setup was slightly different from standard IR. The users saw sets of ten simplified (Wikipedia) documents, about half of which were relevant to a topic given to them beforehand, while the remaining documents were of other randomly selected topics. Based on the gaze pattern, an implicit query was constructed and used to rank unseen documents. The results were significantly better than random rankings.

We extend these results in two ways. First, we will use the eye tracking-based feedback in a more realistic IR scenario. Instead of randomly sampled documents, the user is shown a ranked list of top-5 results, and the task of the system is to improve the ranking of the yet unseen documents. Second, we will improve on the methodology. In (Hardoon et al., 2007) we introduced a two-stage prediction algorithm ("SVM model" in the following), where the latter stage was an SVM which classified new documents into relevant and irrelevant, given a parameter vector that consists of a weight for each word. The parameter vector was inferred with a regressor which had been trained to predict the weight of a word based on the eye movement pattern on the

word. The problem with this method is that learning of the regressor requires a ground truth which is tricky. In the earlier paper we used the parameter vector of an SVM taught to classify the relevant and irrelevant documents in the off-line learning stage, based on their textual content. This is intuitively a sensible strategy, and the experimental results validated it, but the choice is unlikely to be optimal.

In this paper, we introduce a probabilistic model for inferring relevance of the documents. Incorporating both of the two stages, inference of the relevance of a new document and inference of the implicit query, into a single generative model solves rigorously the problem of getting the ground truth and the learning procedure will be optimal for the task, given our modeling assumptions. The method indeed outperforms the earlier one (Hardoon et al., 2007).

Learning of the probabilistic model is related to transfer learning and meta-learning. The implicit query is expressed in the model as latent variables shared within each search task. The central task is to learn an implicit query for a topic unseen in the original training phase, which translates to transfer learning. The eye movements from which the implicit query is inferred can be considered as meta-data for the documents.

We test the method in an experimental scenario designed to closely resemble a real information retrieval setup. The user makes a query within a restricted Wikipedia corpus located in our customized Wikipedia server. A search engine then ranks the top-10 documents for this query, and the first five are shown sequentially to the user using a web browser. The browser has been modified to record and transmit the eye movement measurements to the Wikipedia server. We rank the remaining 5 documents by our eye movement-based model and aggregate the new and the original ranking to produce an ordering for the remaining 5 documents. Average precision in the re-ranked 5 documents was used as the goodness criterion.

Our method, once trained, consists only of a linear discriminator applied to term-specific gaze and term features. Therefore, the method can be applied efficiently in linear time whenever the eye tracking data is available.

2. The Information Retrieval Task

The usual approach in IR is to rank the documents based on their match to a textual query (Baeza-Yates & Ribeiro-Neto, 1999). In our setup the user types in a textual query that reflects her interest but is typically an incomplete description. Then the search engine shows her the top-ranked documents. The eye movements of the user are measured while she reads the documents presented sequentially in the ranked order. Our objective is to use the gaze patterns to improve the ranking of the yet unseen documents. These documents are consequently shown to the user in an order modified using the implicit query inferred from the gaze patterns. In doing so the relevant documents are hopefully shown to the user earlier, that is, the average precision of the search result is improved.

2.1. Okapi BM25 Ranking Function

A widely used ranking function is given by Okapi BM25 (Robertson & Walker, 1994; Robertson & Zaragoza, 2007), which is also used as a baseline method throughout this paper. Okapi BM25 ranks the documents given a textual query q that is a set of terms. Our approach is independent of the actual ranking function, however; indeed, in this work, we could replace Okapi BM25 with any information retrieval system that outputs a ranking of documents for a given query.

Consider a document collection C where each document $d = \{tf_t\}_{t \in V}$ in the collection is a vector of term frequencies, where tf_t is the frequency of term t in the document and V is the vocabulary. For ad hoc retrieval the BM25 weighting function can be expressed as

$$w_t(d, C) = \frac{(1 + k_1) tf_t}{k_1 \left((1 - b) + b \frac{dl}{avdl}\right) + tf_t} \log \frac{|C| - df_t + \frac{1}{2}}{df_t + \frac{1}{2}}, \quad (1)$$

where df_t is the document frequency of term t, dl is the document length and $avdl$ is the average document length across the collection. The k_1 and b are free parameters which we for the purposes of this paper fix to $k_1 = 1.2$ and $b = 0.75$, as suggested by Robertson and Walker (1999). The documents are ranked according to the sum of the weights of the terms in query q:

$$W(d, q, C) = \sum_{t \in q} w_t(d, C). \quad (2)$$

2.2. Metasearch

We have at our disposal a separate and independent ranking system, described in detail in Section 3, that ranks the yet unseen documents based on the gaze patterns of the users. In effect, we have two rankings: the ranking derived from the textual query and given by the Okapi BM25 ranking function described in Section 2.1, and the ranking derived from the gaze patterns.

The problem of combining multiple search engine rankings into one is known as *metasearch* and it has been studied extensively during the past years (see, for example, Cohen et al., 1998; Aslam & Montague, 2001). Because in this work we aim for simplicity and robustness, we use a straightforward linear combination of rankings. Another reason for this choice is that we want our approach to work also with a "black box" search engine which only gives us a ranking of the documents, without any probability of relevance associated with the documents.

In more detail, let $r_{BM25}(d)$ and $r_{EYE}(d)$ be the ranks of the document d given by the Okapi BM25 ranking function and the eye movement model, respectively. We re-rank the yet unseen documents using $score(d)$ defined by

$$score(d) = \gamma r_{BM25}(d) + (1-\gamma) r_{EYE}(d), \quad (3)$$

with the document having the smallest $score(d)$ ranked first. Here γ is a constant between zero and one.

3. Learning to Learn: A Probabilistic Model

In this section we introduce a probabilistic model that can be used to infer the ranking of yet unseen documents for a new and unknown query, given how the user has viewed a set of documents. In practice, the viewed documents are the highest-ranked documents for a given unknown query, and the inferred ranking is used to modify the order in which the further documents are presented.

3.1. Probabilistic Model

The available data is a collection of documents, encoded as TFIDF vectors $\mathbf{d}$, where the component corresponding to term t is

$$\mathbf{d}_t = tf_t \log \frac{|C|}{df_t}.$$

For the viewed documents we additionally have eye movement features. The feature vector $\mathbf{e}_{it}$ contains feature values for term t in document i. There are two types of features: eye movement features that are computed from the eye movement pattern over the term, and textual features that depend only on the term and its location in the document. The features are described in Section 4.3.

In the model, the relevancy r of a document is assumed to depend on the TFIDF vector $\mathbf{d}$ of the document and a search task specific *query vector* $\mathbf{w}$. Our main assumption is that the importance of a certain term for a search query depends only on the way the term is viewed, not on the meaning of the term. This allows us to learn global parameters α and $\boldsymbol{\beta}$, which are common for all search tasks, for the mapping from the features $\mathbf{e}$ to the term's weights w_t. The model is illustrated in Figure 1.

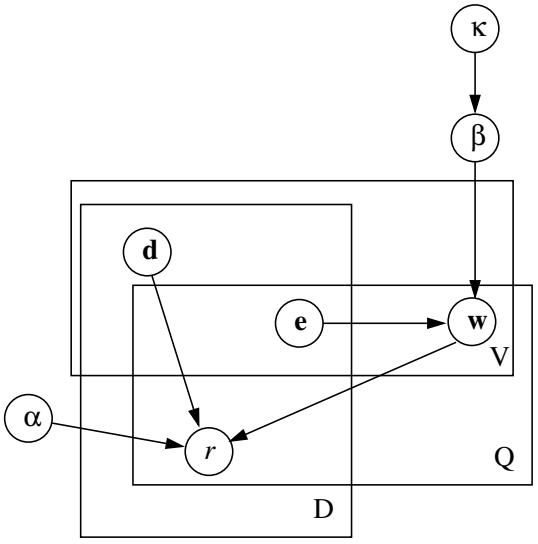

Figure 1. Graphical representation for the generative process for learning to learn. The plates are repeated the number of times shown in their bottom right corner; V is the number of terms in the vocabulary, Q of search tasks, and D of documents in the collection. The $\mathbf{d}$ is the TFIDF representation of the document and r is the 0–1 relevance of a document in a given search task. The $\mathbf{e}$ are the term-specific eye movement and text features and $\mathbf{w}$ is the query inferred from the eye movement features. The α and $\boldsymbol{\beta}$ are parameters shared by all search tasks, and κ is a prior parameter.

The query vector $\mathbf{w}$ of a search task is a vector in $\mathbb{R}^{|V|}$, where $|V|$ is the number of terms in the vocabulary. The entries in the query vector can be interpreted as the relative importance of the terms for the query. We assume that the entries in the query vector are normally distributed, with the mean depending on the eye movement features during the search task. We further assume that the mapping from the eye movements to the query weights is universal in the sense that the parameters of the mapping depend neither on the search task nor on the specific term.

The query weight w_{qt} for term t in the search task q depends on the viewed documents (in our experiments top-k with $k = 5$) in the search task, denoted by D_q, and on all term features in the search task, denoted collectively by $\mathbf{E}^q$. We assume that the dependency is

linear,

$$p(w_{qt}|\mathbf{E}^q, \boldsymbol{\beta}, \boldsymbol{\beta}') = N\left(\frac{1}{|D_q|}\sum_{i\in D_q}\left(\boldsymbol{\beta}^T\mathbf{e}_{it}I_{it} + \boldsymbol{\beta}'^T\mathbf{e}_{it}(1-I_{it})\right), \sigma^2\right), \quad (4)$$

where the indicator variable $I_{it} = 1$ if term t is viewed in document i, and 0 otherwise. If a term appears in the document but is not viewed only the textual features have non-zero values. If a word t does not appear in document i, the term's features are set to zero: $\mathbf{e}_{it} = \mathbf{0}$. The two regression coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$, for viewed and unviewed terms, respectively, are common for all tasks. For notational simplicity we denote both of these parameters by $\boldsymbol{\beta}$ in the following.

For the probability of relevance r of a document in a search task q we assume the functional form of logistic regression. The probability is assumed to be a sigmoidal function of the dot product of the document TFIDF vector $\mathbf{d}$ and the query vector $\mathbf{w}_q$,

$$p(r|\mathbf{d}, \mathbf{w}_q, \alpha) = \frac{1}{1 + e^{-(\alpha + \mathbf{d}^T\mathbf{w}_q)}}. \quad (5)$$

The parameter α is common for all tasks.

We assume availability of a collection of background tasks for learning. Each background task is a search session where we know the relevances of the displayed documents, and have observed users' eye movements. The background tasks are used to learn the shared parameters.

The hyperparameters of the model, α and $\boldsymbol{\beta}$ on which the transfer learning is based, are estimated by maximizing the posterior from which all other parameters have been marginalized out. The logarithm of the marginalized posterior to be maximized is

$$\begin{aligned} L &= \sum_{i\in D_{BG}} \log p(r_i|\mathbf{d}_i, \alpha, \boldsymbol{\beta}) + \log p(\boldsymbol{\beta}|\kappa) \\ &= \sum_{i\in D_{BG}} \log \int p(r_i|\mathbf{d}_i, \mathbf{w}_{q(i)}, \alpha) p(\mathbf{w}_{q(i)}|\mathbf{E}^{q(i)}, \boldsymbol{\beta}) d\mathbf{w}_{q(i)} \\ &\quad + \log p(\boldsymbol{\beta}|\kappa) \\ &\approx \sum_{i\in D_{BG}} \log p(r_i|\mathbf{d}_i, \hat{\mathbf{w}}(q(i), \boldsymbol{\beta}), \alpha) + \log p(\boldsymbol{\beta}|\kappa), \end{aligned}$$

where $q(i)$ is the index of the background task during which document i was shown, D_{BG} is the set of all top-k documents in background tasks, and $\log p(\boldsymbol{\beta}|\kappa) = -\kappa/2\boldsymbol{\beta}^T\boldsymbol{\beta}$ is a Gaussian prior for $\boldsymbol{\beta}$. We assume a uniform prior for α. In the last step, the integral is approximated for computational reasons by a point estimate evaluated at the mode $\hat{\mathbf{w}}$ of $p(\mathbf{w}_{q(i)}|\mathbf{E}^{q(i)}, \boldsymbol{\beta})$; the entries of the mode are

$$\begin{aligned} \hat{w}_t(q, \boldsymbol{\beta}) &= \arg\max_{w_{qt}} p(w_{qt}|\mathbf{E}^q, \boldsymbol{\beta}) \\ &= \frac{1}{|D_q|}\sum_{i\in D_q}\left(\boldsymbol{\beta}^T\mathbf{e}_{it}I_{it} + \boldsymbol{\beta}'^T\mathbf{e}_{it}(1-I_{it})\right). \quad (6) \end{aligned}$$

Here $|D_q| = k$ is the number of viewed documents (top-k documents) in the search task q.

The learned values of the hyperparameters α and $\boldsymbol{\beta}$ are used to transfer knowledge from the old tasks to a new one. For the new task, we observe only eye movements on a small number of documents; we do not know the relevances of the documents as in the learning phase. The best prediction of relevance would result by integrating over the potential query vectors, but for computational reasons we again estimate the integral by the mode of $p(\mathbf{w}_{q(i)}|\mathbf{E}^{q(i)}, \boldsymbol{\beta})$. The mode can be interpreted as the estimated query vector $\mathbf{w}_{new}$, estimated using the equation (6), where the sum now is over the documents in the new task.

The ultimate goal is to find documents which are relevant to the new query. We rank the test set of unseen documents according to the probabilities (5) computed using $\mathbf{w}_{new}$.

3.2. SVM Model

The probabilistic model was motivated by the model (denoted by "SVM model") of Hardoon et al. (2007); we compared our model with the linear variant of the SVM model that performed almost as well as the best one in the original paper.

The main difference between our probabilistic model and the SVM model is that we have a full generative framework for all observations. A useful consequence of this is that we have a principled way for generating the search task specific implicit query $\mathbf{w}$ from the term-specific eye movements patterns. In the SVM model this step was somewhat ad hoc; the ground truth was obtained by classifying relevant vs. irrelevant documents in the training data. The regressor then tried to predict the SVM discriminator weights from eye tracking data. There is no guarantee that the discriminator weights used by the SVM are optimal, or even always good, targets taking into account uncertainties stemming from the noisy eye movements.

Note that in this paper we used fairly simple computational approximations for generating the task-specific implicit queries $\mathbf{w}$. The fact that the results are still good gives the model further support; if necessary,

more accurate approximations can be developed later.

3.3. Connection to Transfer Learning and Meta-Learning

In our problem we need to learn to learn an implicit query from the text of the document and gaze pattern. This needs to be done for search topics unseen in the training phase. Each search topic has a hidden representation that we have to learn, namely the query vector $\mathbf{w}$. We have additionally introduced hyperparameters, namely the α and $\boldsymbol{\beta}$, that are shared across all search tasks. The shared parameters contain information that is needed to learn the search task-specific query vector.

Our modeling assumption is that there exists information in the gaze pattern that is independent of the actual semantic content of the words and of the specific query. We encode independence of the specific query by introducing the parameters α and $\boldsymbol{\beta}$ that are shared across the search tasks. Independence of the semantic content is achieved by constructing the model so that the query vector $\mathbf{w}$ depends only on text and eye movement features associated with a specific term, but not on the semantic content of the terms. In other words, the model is invariant with respect to any permutation of term labels.

The learning process, described in Section 3.1, can be interpreted to have two phases: in the first phase the parameters α and $\boldsymbol{\beta}$ that are shared across all search tasks are learned using several background search tasks. In the second "on-line" phase a search task-specific query vector $\mathbf{w}$ (for a previously unseen search task) is estimated using the shared parameters.

Our problem is related to transfer learning and meta-learning (Thrun, 1996; Baxter, 2004; Caruana, 1997; Ando & Zhang, 2005; Thrun, 1998; Pratt & Thrun, 1997; Vilalta & Drissi, 2002; Giraud-Carrier et al., 2004). Transfer learning and meta learning utilize data from other "similar" learning tasks and from multiple applications of the learning system. For example, learning to recognize objects in cartoons might help to recognize objects in photographs (Elidan et al., 2006); or in our case, learning to infer query vectors in search tasks helps in learning a query vector in a yet unseen search task. We can think that in our case the inductive bias extracted is parametrized by α and $\boldsymbol{\beta}$ and fixed when these parameters are learned. When we observe a new search task we can then use the information coded in α and $\boldsymbol{\beta}$ to learn the task-specific query vector $\mathbf{w}$ that can finally be used to predict the relevance of a given document.

4. Experiments

We conducted small scale eye tracking experiments to validate the proposed model. The experiments were designed to simulate the common case where some keywords are available but they are not a sufficient description of the interests.

4.1. Search Tasks

We constructed 13 search tasks for text documents (see Table 1). The search tasks were chosen prior to doing any experiments. The criteria for selecting the search tasks were that there should be several relevant documents for each task and, furthermore, that the original query should also suggest irrelevant documents. That is, there should be irrelevant documents which are ranked quite high. This is why the search terms were purposefully ambiguous.

For example, in search task number 3 the task was to find information about "ancient Rome." The user would be instructed to find documents that would tell about ancient Rome. The textual search query, forced by us, was "Rome." As a result, the search results included articles also, for example, of modern Rome. Our purpose was to see whether the gaze pattern could be used to infer a new query. Intuitively, the query vector $\mathbf{w}$ inferred from the eye movements could include with positive weight terms related to ancient Rome, such as "ancient", "Caesar" and "Carthage"; and possibly with negative weight terms related to the modern times, such as "airport" or "president".

The document corpus consisted of articles downloaded from Wikipedia. Only the lead section of the documents before the first section header was shown to the user and used in the experiments. For each search task, we selected 10 documents having the highest BM25 score.

4.2. Experimental Procedure

There were three participants in the experiments, one female and two males. The test subjects were voluntary under and post-graduate researchers (the authors were not included).

The participants were asked to search for documents of a given topic using a web search engine. They were advised to act as if they were collecting the relevant documents of the topic for later reading, that is, they were supposed to stop reading the document once they had determined whether the document was relevant. Each task was started by clicking a search button which submitted the pre-entered search term to a custom server.

Table 1. The search tasks and the given search terms.

Task number	Desired topic	Search term
1	American football	football
2	Alternative medicine	medicine
3	Ancient Rome	Rome
4	Adhesive tape	tape
5	Environmental conservation	conservation
6	Seal (marine mammal)	seal
7	Extra-solar planets	planets
8	Visual nervous system	vision
9	Internet forums	forum
10	Marketing strategies	strategies
11	National libraries	libraries
12	British Royal Navy	navy
13	Space shuttles	shuttle

Our search engine did not return a list of the most relevant documents as search engines usually do. Instead, it returned the most relevant document directly. In the bottom of each document there were two links which were used for marking the document as either relevant or not relevant. Clicking one of these links retrieved the next document. All other links were removed from the documents. The search engine returned the documents in the order determined by the BM25 algorithm. Each search session included 10 documents. After finishing one session the test subject was automatically given a topic and search terms for the next task.

The relevance judgements given by the users were used as ground truth during the training phase of the model. In testing phase they were used to validate the results.

During the search tasks the users' eye movements were recorded with a Tobii 1750 eye tracker. Tobii tracks gaze location by measuring the reflection pattern on the cornea of eye. It does not require wearing a helmet or a headrest. The users were sitting 60 cm away from a 17 inch computer screen. The system was calibrated once in the beginning of the experiment.

4.3. Term Features

The eye tracker and the browser recorded the sequence of fixations; it was then transmitted to the Wikipedia document server when the user clicked any link. A part of the gaze trajectory was considered a fixation if the gaze stayed inside a 30 pixel square (about 0.6 visual angle) for more than 100 ms. Fixations that appeared outside the bounding boxes of vocabulary words were ignored. The vocabulary consisted of stemmed words, with stop words removed. The size of the vocabulary was 3030 words.

For each term t, we extracted 19 eye movement features and 3 text features, denoted collectively by $\mathbf{e}_t$. We used the same eye movement features, such as number of fixations, absolute, relative and mean fixation durations, used by Hardoon et al. (2007) as well. The 3 text features were independent of the actual search task; they are the number of characters in the word, the relative position of the word in the document, and the inverse document frequency of the word.

4.4. Combination of Textual and Eye Movement Based Searches

We show that a simple combination of our eye movement-based ranking and a ranking by a state-of-the-art textual IR algorithm has higher precision than the textual search alone. For the textual search we use BM25, a well-known bag-of-words ranking function.

We measured the eye movements while the users were reading top-5 documents as returned by BM25. The documents below the rank 5 were used for testing. We ranked the test documents with our method thus getting a second ranking in addition to the original BM25 ranking. We combined the two rankings by reordering the documents according to the weighted average (3).

To compare the original BM25 ranking and the combined ranking, we compare their average precision, a common measure for evaluating search results. It is computed as the average of the precisions at positive rankings: $\frac{1}{R}\sum_{i=1}^{R}\frac{i}{r_i}$, where R is the total number of relevant documents, and the r_i are the rankings of the positive documents such that $r_i < r_{i+1}$. The best possible average precision is one, which corresponds to all relevant documents being ranked in the first positions.

We learn the query vector for each search task by leaving the data for that task out and using the remaining tasks as background tasks in the first phase of the training, as was discussed in Section 3.1. In the "online" phase we use the top-5 documents from the left-out task to infer the query vector.

We combine BM25 and the probabilistic model by using equation (3). For that purpose we need to decide a value for the weighting factor γ. We compare the mean average precision of the plain BM25 and the combination of BM25 and the proposed probabilistic model for documents that are ranked 6–10 in each background task for several discrete values of γ, and for each task select the γ value that has the best mean improvement in average precision. We use the average of these task-specific best values of γ in order to obtain a common value of $\gamma = 0.2$, which is then the value used in all of the experiments. The value $\kappa = 1$ for the prior pa-

rameter is selected in an analogous fashion at the same time.

The results are shown in Table 2. On the test set, in 9 search tasks out of 13 the mean average precision of the combined ranking across test subjects outperformed the baseline BM25 ranking. The difference in average precision is statistically significant ($p = 0.047$, one-tailed Wilcoxon Signed Rank Test).

The worst performance is shown by task 6 in which the only relevant document in ranks 6–10 (according to the users' relevance judgement) was moved from being 6th to 10th, thus reducing the average precision for documents in ranks 6–10 from unity of the BM25 baseline down to 0.29. In all the other tasks the performance either improved clearly, or for some degraded slightly.

In order to examine the relative contributions of the eye movement and textual features we compared two probabilistic models, one using just the text features (of all words, irrespective of whether they were looked at or not) and one using all the features. The mean average precision of the latter was 6.3 percentage units higher but for this amount of data the difference was not statistically significant ($p = 0.22$).

4.5. Comparison to the SVM Model

We compared the performance of the combination of BM25 and the probabilistic model to an analogous combination of BM25 and the SVM (of Section 3.2). The SVM model is trained using the eye movements on the training documents, as described in (Hardoon et al., 2007), and the resulting ranking is combined to the BM25 ranking identically as was done above for the probabilistic model.

The mean average precision of the combination of BM25 and the SVM is worse than the average precision of BM25 for all values of the weighting parameter γ. The bad performance of the SVM model here is probably due to the fact that the ground truth for the query vectors, estimated with SVM from the very small data sets, is likely to be very noisy. It is also possible that the probabilistic model is otherwise better suited for the relatively small training set sizes.

5. Discussion

We introduced a generative model of how the relevance of documents is related to the viewing patterns of people during a search task. The model is trained in two phases. In the first phase, the hyperparameters that are independent of the search topic are learned. In the second "on-line" phase, the learned parameters are used to infer an implicit query from the gaze patterns during a new search session.

The system is realistically applicable; we have indeed implemented it using a standard web browser that has been modified to record and transmit the information about the gaze patterns to the web server. The document corpus used in the experiments consisted of abstracts of Wikipedia articles.

Our results imply that the performance of the method correlates with the search task; in particular, there are some tasks for which the methods seems to perform quite badly although on average it clearly improves the results. It needs to be investigated more carefully later, which kinds of search tasks the eye movements are helpful in, and whether different types of models are useful for different kinds of tasks.

Acknowledgements

KP belongs to the Finnish Centre of Excellence in Algorithmic Data Analysis, and AA and SK to Finnish Centre of Excellence in Adaptive Informatics Research. We would like to thank David Hardoon and John Shawe-Taylor for valuable input, Wray Buntine for the Wikipedia data and Janne Kataja for the Gazillion browser. This work was supported in part by the PASCAL2 Network of Excellence of the European Community.

References

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Aslam, J. A., & Montague, M. (2001). Models for metasearch. *SIGIR '01: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 276–284). New York, NY: ACM.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Reading, MA: Addison-Wesley.

Baxter, J. (2004). A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, *28*, 7–39.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. *IUI '01: Proceedings*

Table 2. Mean improvement in the average precision between plain BM25 and the combination of BM25 and the proposed probabilistic model (top row, $\Delta AVGPREC_{Prob}$). The mean average precision of the plain BM25 which shows the baseline performance on purely textual searches is shown in the bottom row, titled $AVGPREC_{BM25}$. The probabilistic model outperforms the BM25 baseline model ($p = 0.047$, one-tailed Wilcoxon Signed Rank Test), while the performance of the SVM model is comparable to the baseline. Larger values are better.

	Search task												
	1	2	3	4	5	6	7	8	9	10	11	12	13
$\Delta AVGPREC_{Prob}$	0.17	0.28	0.53	0.18	0.22	-0.71	-0.03	0.16	0.04	0.14	-0.05	-0.08	0.34
$AVGPREC_{BM25}$	0.50	0.61	0.13	0.50	0.50	1.00	0.91	0.50	0.78	0.70	0.71	0.50	0.42

of the International Conference on Intelligent User Interfaces (pp. 33–40). New York, NY: ACM Press.

Cohen, W. W., Schapire, R. E., & Singer, Y. (1998). Learning to order things. *NIPS '97: Proceedings of the Conference on Advances in Neural Information Processing Systems 10* (pp. 451–457). Cambridge, MA: MIT Press.

Elidan, G., Heitz, G., & Koller, D. (2006). Learning object shape: From drawings to images. *CVPR'06: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2*, 2064–2071.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, *23*, 147–168.

Giraud-Carrier, C., Vilalta, R., & Brazdil, P. (2004). Special issue on meta-learning. *Machine Learning*, *54*, 187–312.

Hardoon, D. R., Ajanki, A., Puolamäki, K., Shawe-Taylor, J., & Kaski, S. (2007). Information retrieval by inferring implicit queries from eye movements. *AISTATS '07: The International Conference on Artificial Intelligence and Statistics*. Electronic proceedings at www.stat.umn.edu/∼aistat/proceedings/start.htm.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *SIGIR '05: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 154–161). New York, NY: ACM.

Joachims, T., & Radlinski, F. (2007). Search engines that learn from implicit feedback. *IEEE Computer*, *40*, 34–40.

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

Pratt, L., & Thrun, S. (1997). Second special issue on inductive transfer. *Machine Learning*, *28*, 5–130.

Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., & Kaski, S. (2005). Combining eye movements and collaborative filtering for proactive information retrieval. *SIGIR '05: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 146–153). New York, NY: ACM.

Robertson, S., & Zaragoza, H. (2007). The probabilistic relevance model: BM25 and beyond. Tutorial at the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), http://www.sigir2007.org/tutorial2d.html.

Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *SIGIR '94: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 232–241). New York, NY: ACM.

Robertson, S. E., & Walker, S. (1999). Okapi/Keenbow at TREC-8. *Proceedings of the Eighth Text Retrieval Conference TREC-8* (pp. 151–162). Washington, DC: GPO.

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? *In Advances in Neural Information Processing Systems 8* (pp. 640–646). Cambridge, MA: MIT Press.

Thrun, S. (1998). *Learning to learn*, chapter Lifelong learning algorithms, 181–209. Norwell, MA: Kluwer Academic Publishers.

Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. *SIGIR '06: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 11–18). New York, NY: ACM.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, *18*, 77–95.