Implicit relevance feedback from eye movements

Jarkko Salojärvi[1], Kai Puolamäki[1], and Samuel Kaski[1,2]

[1] Neural Networks Research Centre, Helsinki University of Technology
P.O. Box 5400, FI-02015 HUT, Finland
{forename.surname}@hut.fi
[2] Department of Computer Science, University of Helsinki
P.O. Box 68, FI-00014 University of Helsinki, Finland
{forename.surname}@cs.helsinki.fi

Abstract. We explore the use of eye movements as a source of implicit relevance feedback information. We construct a controlled information retrieval experiment where the relevance of each text is known, and test usefulness of implicit relevance feedback with it. If perceived relevance of a text can be predicted from eye movements, eye movement signal must contain information on the relevance. The result is that relevance can be predicted to a considerable extent with discriminative hidden Markov models, and clearly better than randomly already with simple linear models of time-averaged data.

1 Introduction

A search engine could be improved by an algorithm which models the interests of a user. Such an algorithm would be *proactive*; it would predict the needs of the user and adapt its own behavior accordingly [1]. The usual way to learn the interests of the user would be to ask, after each document, whether the user found it relevant, and to learn the user's preferences from the answers. However, giving this kind of explicit feedback is laborious.

Alternatively, relevance can be inferred from implicit feedback derived traditionally from document reading time, or by monitoring other behavior of the user (such as saving, printing, or selecting of documents). The problem with the traditional sources is that the number of feedback events is relatively small.

In this paper we explore whether the traditional sources of implicit relevance information could be complemented with eye movements. We construct an experimental information retrieval setup where relevance of the texts is known. The measured eye movement data corresponding to each text will then have a known label, and the data can be used as a learning data set for machine learning methods. Machine learning will be used for selecting a good set of features and for learning time series models to predict relevance of new measurements.

We make an assumption that human attention patterns correlate with relevance; at the simplest, people tend to pay more attention to objects they find relevant or interesting. The reason why gaze direction provides an indicator of the focus of attention lies in the physiology of the eye. Accurate viewing is possible only in the central *fovea* area (only 1–2 degrees of visual angle) where the

density of photoreceptive cells is highly concentrated. A scene needs therefore to be inspected with a sequence of alternating *saccades* (rapid eye movements) and *fixations* (the eye is fairly motionless). Information on the environment is mostly gathered during fixations, and fixation duration is known to be correlated with the complexity of the object under inspection. A simple physiological reason for this is that the amount of information the visual system is capable of processing is limited. In other words, we assume that (visual) attention lies there where the amount of gathered information is larger. This is justified in a multitude of psychological experiments [2].

Related Work. Eye movements have traditionally been exploited as an alternative input modality in user interfaces, for instance in eye typing (cf. [3]). Another increasingly popular application area is usability studies, where the goodness of a user interface or a web page is evaluated by monitoring natural behavior of the users. This form of implicit feedback information gathered from eye movements has been analyzed using features computed for larger areas of interest [4], such as images or captions of text.

As far as we know, eye movements have been used in information retrieval applications in only two previous studies. The first is our earlier preliminary work [5, 6], which is extended in this paper by thorough experiments with a large number of subjects, better equipment that solves our earlier calibration problems, and more detailed analysis of the results. The goal of the second related work [7] was different: to investigate with quantitative measures how users behave in a real, less-controlled information retrieval task.

2 Information retrieval experiment

In a typical information retrieval setup the user types in keywords to a search engine and is then given a list of results, for instance titles of scientific articles, that possibly contain the information the user is looking for. Some of the proposed titles may be totally irrelevant, some of them handle the correct topic, and only few are links to articles that the user eventually reads. Our experimental setting for collecting eye movement data was designed to simulate this natural situation, with the difference that in our case the relevance is known.

The subject was first shown a question, and then a list of ten sentences, one of which contained the correct answer (C). Five of the sentences were known to be irrelevant (I), and four relevant to the question (R). The task of the subject was to identify the correct answer, press 'enter' (this ended the eye movement measurement), and then type in the associated number in the following display. Each of the eleven test subjects carried out 50 assignments. The assignments were in Finnish, the mother tongue of the subjects. Eye movements were measured with a Tobii 1750 eye tracker with a screen resolution of 1280x1024 and a sampling rate of 50 Hz.

Preprocessing. The raw eye movement data (x and y coordinates of the gaze direction) was segmented into a sequence of fixations and saccades by a window-based algorithm (software from Tobii).[3] Each fixation was first assigned to the nearest word. A set of 22 features were computed from the eye movement trajectory for each word [10]; similar features are used in psychological studies of reading [2]. The resulting feature vector sequence was then segmented to subsequences corresponding to different sentences, and a label was assigned to each subsequence according to the known class of the sentence.

Sentence-specific averages of the features was then computed. Linear Discriminant Analysis (LDA) was applied to the averaged features in order to select the set of features that best predict relevance for left-out data. The LDA also provided a baseline classification accuracy.

In the time-series models, described in more detail below, the resulting set of features were modeled with the following exponential family distributions: (1) One or many fixations within the word (binomial). (2) Logarithm of total fixation duration on the word (assumed Gaussian). (3) Reading behavior (multinomial): skip next word, go back to already read words, read next word, jump to an unread line, or last fixation in an assignment.[4]

3 Models

The simplest method to classify the eye movement data is to disregard the time dependency between data samples and compute averages of the eye movement features. This gives sentence-specific feature vectors, one per sentence. In this paper, the averaged vectors were classified with Linear Discriminant Analysis. More fine-grained cues of relevance can be sought with models that take into account the time series nature of the eye movement data.

Hidden Markov Models. The simplest model that takes the sequential nature of data into account is a two-state hidden Markov model (HMM). A separate model was optimized individually for each class, by maximizing the log-likelihood of the data Y given the model and its parameters Θ, that is, $\log p(Y|\Theta)$. The HMMs are trained with the Baum-Welch (BW) algorithm [8]. In a prediction task the models of different classes were combined into a maximum a posteriori (MAP) prediction.

Discriminative Hidden Markov Models. In speech recognition, where HMMs have been extensively used for decades, the current state-of-the-art HMMs are discriminative. Discriminative models aim to predict the relevance $B = \{I, R, C\}$

[3] 20-pixel window size and a minimum duration of 80 ms.

[4] Feature selection was carried out with methods that use averaged data (from eigenvectors of LDA), since currently there are no methods that can simultaneously both do this and model the time series. We therefore chose to model a representative set of features which can be used to construct the best discriminating averaged measures.

of a sentence, given the observed eye movements Y. Formally, we optimize $\log p(B|Y, \Theta)$. In discriminative HMMs, a set of states or a certain sequence of states is associated with each class. This specific state sequence then gives the probability of the class, and the likelihood is maximized for the teaching data, versus all the other possible state sequences in the model [9]. The parameters of the discriminative HMM can be optimized with an extended Baum-Welch (EBW) algorithm, which is a modification of the original BW algorithm.

We model eye movements with a two-level discriminative HMM, where the first level models transitions between sentences, and the second level transitions between words within a sentence. Viterbi approximation is used to find the most likely path through the second level model (transitions between words in a sentence), and then the discriminative Extended Baum-Welch optimizes the full model.

Voting. The HMMs produce probabilities for the relevance classes (I, R, C) for each viewed sentence. However, the users may look at a sentence several times, and the resulting probabilities need be combined in a process we call *voting*.

We constructed a log-linear model for combining the predictions. Assume that the sentence-specific probability distribution, $p(B|Y_{1\ldots K})$, can be constructed from the probability distributions of the kth viewings of the sentence, $P(B|Y_k)$ (obtained as an output from a Markov model) as a weighted geometric average, $p(B|Y_{1\ldots K}, \alpha) = Z^{-1} \prod_k p(B|Y_k)^{\alpha_{Bk}}$, where Z is a sentence-specific normalization factor and the parameters α_{Bk} are non-negative real numbers, found by optimizing the prediction for the training data. The predicted relevance of a sentence is then the largest of $p(I)$, $p(R)$ and $p(C)$.

It is also possible to derive a simple heuristic rule for classification by assuming that decision of relevance is made only once while reading the sentence. We will call this rule maxClass, since for each sequence we will select the maximum of the predicted relevance classes (with ordering $I < R < C$). A simple baseline for the voting schemes is provided by classifying all the sequences separately (i.e., no voting).

4 Results

The prediction accuracy was assessed with 50-fold cross validation, in which each of the assignments was in turn used as a test data set. In order to test how the method would generalize to new subjects, we also ran an 11-fold cross validation where each of the subjects was in turn left out. Table 1 lists the classification accuracies, that is, the fraction of the viewed sentences in the test data sets for which the prediction was correct. The methods generalize roughly equally well both to new assignments and to new subjects. The performance of the two different voting methods (log-linear and maxClass) seem to be nearly equal, with log-linear voting having a slight advantage.

Table 2 shows the confusion matrix of the discriminative HMMs. Correct answers (C) are separated rather efficiently. Most errors result from misclassifying

Table 1. Prediction accuracies of different models. The baseline is given by the the "dumb model," which classifies all sentences to the largest class I. Differences between LDA and dumb classifier, and simple HMMs and LDA, tested significant ($P < 0.01$, McNemar's test), as well as the difference between discriminative HMM and simple HMMs in the case of leave-one-assignment-out validation (with log-linear voting). Left column: obtained by 50-fold cross-validation where each of the assignments was left out in turn as test data. Right column: Obtained by 11-fold cross-validation where each of the subjects was in turn left out to be used as test data.

Method	**Accuracy (%)** (leave-one-assignment-out)	**Accuracy (%)** (leave-one-subject-out)
Dumb	47.8	47.8
LDA	59.8	57.9
simple HMMs(no vote)	55.6	55.7
simple HMMs(maxClass)	**63.5**	**63.3**
simple HMMs(loglin)	**64.0**	**63.4**
discriminative HMM(loglin)	**65.8**	**64.1**

relevant sentences (R) as irrelevant (I). It is also possible to compute precision and recall measures commonly used in information retrieval by treating correct answers as the relevant documents. The resulting precision rate is 90.1 %, and recall rate 92.2 %.

Table 2. Confusion matrix showing the number of sentences classified by the discriminative HMM, using loglinear voting, into the three classes (columns) versus their true relevance (rows). Cross validation was carried out over assignments. The percentages (in parentheses) denote row- and column-wise classification accuracies.

	Prediction		
	I (62.4 %)	R (61.8 %)	C (90.1 %)
I (77.3 %)	1432	395	25
R (43.6 %)	845	672	24
C (92.2 %)	17	21	447

5 Conclusion

Our results show that relevance information can be inferred from eye movement signals. Both the time series nature of the data and the discriminative nature of the task should be taken into account when constructing models for eye movements. An interesting topic for further research would be to inspect whether the HMM is capable of differentiating cognitive processes associated with reading.

The work will be continued in the form of a PASCAL eye movement challenge [10]. This will hopefully result in a toolbox of robust and efficient methods for relevance extraction.

6 ACKNOWLEDGMENTS

The authors thank Jaana Simola for carrying out the eye movement measurements, and Lauri Kovanen for implementing part of the code. We also thank Maria Sääksjärvi and Johanna Gummerus, Center for Relationship Marketing and Service Management at Swedish School of Economics and Business Administration (Hanken), for allowing us to use their eye tracker. This work was supported in part by Academy of Finland, decision 79017, and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. All rights are reserved because of other commitments.

References

1. Tennenhouse, D.: Proactive computing. Commun. ACM **43** (2000) 43–50
2. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin **124** (1998) 372–422
3. Ward, D.J., MacKay, D.J.: Fast hands-free writing by gaze direction. Nature **418** (2002) 838
4. Goldberg, J.H., Stimson, M.J., Lewenstein, M., Scott, N., Wichansky, A.M.: Eye tracking in web search tasks: design implications. In: ETRA '02: Proceedings of the symposium on Eye tracking research & applications, ACM Press (2002) 51–58
5. Salojärvi, J., Kojo, I., Simola, J., Kaski, S.: Can relevance be inferred from eye movements in information retrieval? In: Proceedings of WSOM'03, Workshop on Self-Organizing Maps. Kyushu Institute of Technology, Kitakyushu, Japan (2003) 261–266
6. Salojärvi, J., Puolamäki, K., Kaski, S.: Relevance feedback from eye movements for proactive information retrieval. In Heikkilä, J., Pietikäinen, M., Silvén, O., eds.: workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004), Oulu, Finland (2004)
7. Granka, L.A., Joachims, T., Gay, G.: Eye-tracking analysis of user behavior in WWW search. In: Proceedings of SIGIR'04. ACM Press (2004) 478–479
8. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77** (1989) 257–286
9. Povey, D., Woodland, P., Gales, M.: Discriminative MAP for acoustic model adaptation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). Volume 1. (2003) 312–315
10. Salojärvi, J, Puolamäki, K., Simola, J., Kovanen, L., Kojo, I., Kaski, S.: Inferring relevance from eye movements: feature extraction. Helsinki University of Technology, Publications in Computer and Information Science, Report A82 (2005). `http://www.cis.hut.fi/eyechallenge2005/`