Can relevance be inferred from eye movements in information retrieval?

Jarkko Salojärvi†, Ilpo Kojo‡, Jaana Simola‡, and Samuel Kaski†
† Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 9800, FIN-02015 HUT, Finland
‡ Center for Knowledge and Innovation Research, Helsinki School of Economics, P.O. Box 1210, FIN-00101 Helsinki, Finland
e-mail: {jarkko.salojarvi,samuel.kaski}@hut.fi, {ilpo.kojo,jaana.simola}@hkkk.fi

Keywords: Data exploration, eye movement, information retrieval, learning metrics, relevance feedback

Abstract— We investigate whether it is possible to infer from implicit feedback what is relevant for a user in an information retrieval task. Eye movement signals are measured; they are very noisy but potentially contain rich hints about the current state and focus of attention of the user. In the experimental setting relevance is controlled by giving the user a specific search task, and the modeling goal is to predict from eye movements which of the given titles are relevant. We extract a set of standard features from the signal, and explore the data with statistical information visualization methods including standard self-organizing maps (SOMs) and SOMs that learn metrics. Relevance of document titles to the processing task can be predicted with reasonable accuracy from only a few features, whereas prediction of relevance of specific words will require new features and methods.

1 Introduction

In *proactive information retrieval*, the system adapts to the interests of the user that are inferred from either explicit or implicit feedback. Explicit feedback by indicating which documents are relevant to the query is naturally more accurate but the users often consider it too laborious and time-consuming. The usability and accuracy of information retrieval applications would be greatly enhanced by complementing explicit feedback with implicit feedback signals measured from the user and the interface. Research on implicit feedback potentially has even wider-ranging implications. If it is reliable enough, it will be useful in a range of other applications as well. Ultimately, a genuine personal assistant could adapt to the goals and interests of the user and learn to disambiguate her vague commands and anticipate her actions.

In this paper, we try to infer relevance of read documents based on eye movements measured during reading. Eye movements contain potentially lots of useful information about the user and her interests. The problem is that it is hard to extract the task-relevant characteristics from among all the other variation in eye movements, for example measurement noise, calibration errors, and personal reading and searching strategies.

While it is hard to define interestingness or relevance in general, we construct a controlled experimental setting in which it is known which documents are relevant, and try to learn relevance from data. The user is instructed to find an answer to a specific question, and some of the given document titles are known to be relevant.

In this first feasibility study we extract a set of standard features from eye movements for each title and word, and correlate the features to the known relevance. The two goals are (i) to find out whether relevance can be estimated based on these features, and (ii) which features are important in making the prediction. The data is explored with standard unsupervised information visualization methods (linear principal component analysis and non-linear self-organizing maps [7]), and corresponding discriminative methods (linear discriminant analysis and self-organizing maps that learn metrics [6, 9]).

This paper serves additionally as a case study for the learning metrics principle that has been developed for precisely this kind of tasks. Data needs to be explored since it is not known a priori which dimensions of the primary data (here the eye movements) are relevant. It is known, however, that variation in the primary data is relevant to the extent that it correlates with certain auxiliary data, here an indicator of whether the title is relevant or not.

2 Eye movements: background

Psychologists have studied eye movements as an indicator of different cognitive processes for decades [10]. However, few engineering applications exploiting eye movements have been introduced. Most of them have used eye movements as a controlling interface (for example for eye typing [13]). This kind of approach may easily lead to a problem known as 'Midas touch'; the user needs to control her eye movements consciously in order to avoid initiating unintended commands. However, as a source of implicit feedback, eye movements are a self-evident candidate since gaze is by far one of the most important nonverbal signs of human attention.

2.1 The eye

The neuroanatomical basis for the importance of gaze direction lies in the structure of the retina and the visual pathway. Image is converted into neural signals in the retina and subsequently processed by the brain. Retinal sampling density is uneven: the density is high in the central retina (fovea, only 1-2 degrees of visual angle) and decreases rapidly towards the periphery. In the primary visual cortex the processing of the input from fovea has been alloted more space than processing of the peripheral retina. Functionally, this means that with a single glimpse the human vision can obtain information with high resolution only from a very small central area of the visual field. As a result, gaze direction provides a good indicator of the target of subject's attention in intensive visual tasks such as reading, visual search, or working with computers.

Foveal vision and serial scrutiny of stimuli are needed when analyzing objects with small details such as texts, numbers, and parts of diagrams. The main components of this serial processing are *fixations* (stops) and *saccades* (rapid jumps) of the eye. The retinal image is transmitted to the brain during fixations but not during saccades. It is also evident that when the object under scrutiny (e.g. a word) is complex, the duration of fixation is longer than when the object is simple [10].

2.2 Eye movements and reading

In a typical reading situation, the reader fixates on each word sequentially. Some of the words are skipped, some fixated twice and some trigger a *regression* to preceding words (approx. 15 % of the saccades). The reader is often not conscious of these regressions. The typical duration of fixations varies between 60-500 ms, being 250 ms on the average [8].

In psychological literature, several models for reading have been proposed. Almost all concentrate on modeling reading at the basic level, as a series of sequential fixations occurring from left to right without regressions which are generally assumed to be associated with higher order cognitive processes. The durations of the fixations are correlated with word occurrence frequency [11], that is, the access time for the concepts concerning more rarely occurring words is longer than the access time for more frequently occurring words (however, similar correlations with word predictability and word length have also been reported). In a more recent publication [3] this correlation is extended to explain also regressions as occurring to those words which did not receive enough processing time during the first pass reading.

3 Eye movements and relevance

Few attempts to infer different higher order cognitive processes (such as relevance determination) from eye movements can be found in literature. An interesting publication is [12], where different cognitive processes are modelled with Hidden Markov Models in a simple equation-solving task. A setup most similar to relevance determination is introduced in [4], where a translator is activated if the reader encounters a word which she has difficulties (these are inferred from eye movements) in understanding.

In order to determine relevance from eye movements in an information retrieval task, we devised an experimental setup where the relevant items are known, and then measured the eye movements of test subjects during the experiment.

3.1 Experimental setup

Subjects' eye movements were recorded with a head-mounted gaze tracking system (iView from SensoMotoric Instruments GmbH, Germany). It has two small video cameras; one of them monitored the scene in front of the subjects' eyes and the other one, a small infrared-sensitive camera, monitored the dominant eye while an infrared LED illuminated the eye. Video images of the pupil and corneal reflections of the infrared LED were captured at 50 Hz by the eye tracker, and the gaze points and pupil diameter were computed from that data. Calibration of the eye movement system (using a set of nine screen locations) was carried out several times during an experiment. The subject's head was stabilized with a chin rest to improve the accuracy of calibration and measurements. After calibration the participant was asked not to move.

Three subjects participated in the experiments. Each experiment consisted of twenty sets, each containing two displays; in the first one a task assignment and twelve titles were shown, in the second one numbers for identification were added to the screen. The subjects task was to (i) read the assignment and decide whether there was a title containing the answer to the task assignment, and (ii) tell the number of that title. Eye movements were recorded during reading the first display (i). From the twelve titles, four (including the one containing an answer) were relevant, i.e. dealt with the same topic as the question, and eight were non-relevant. For control purposes some of the assignments did not have a title containing the answer. The experiments were carried out in Finnish, the mother tongue of the subjects. Newspaper headlines were used as titles.

3.2 Common measures from eye movements during reading

From a psychological perspective, the experimental setup described above measures browsing, an intermediate between reading and visual search. Psychological research has mainly focused on visual search and reading, for which several different measures have been devised [2, 10]; the validity of these features for browsing is yet to be determined. We chose a comprehensive set of reading measures in order to find out which of them would be feasible for determining relevance in the experimental setup (see the Appendix for brief descriptions of the measures).

We additionally implemented three new features: Mean pupil width and standard deviation during the reading of a word (since pupil size correlates with workload [1]), and the first fixation duration within a word divided by the total duration of fixations in each set (i.e. the proportion of total processing time a word receives when first encountered). Pupil size changes quite slowly with a delay of approximately one second; hence an effect could possibly be discernible in the level of sentences.

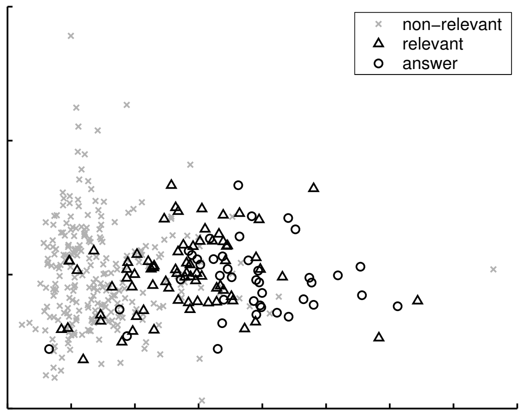

Figure 1: Projection of eye movement data of all subjects to two principal components. *Non-relevant* titles were from a different topic, *relevant* from the correct topic but did not contain an answer to the posed question, *answer* were relevant and contained the answer.

4 Data Analysis

Since hardly anything was known about the data, we began by exploration, "looking at the data." We first applied unsupervised methods to learn about correlations between the variables and cluster structures of the data. Classical principal component analysis and self-organizing maps in the normal Euclidean metric were applied. In the second stage we used discriminative methods to visualize how well the relevant and irrelevant titles can be distinguished, and which features are required for the discrimination. We used here as well a linear method, the classical linear discriminant analysis (LDA), and a new nonlinear method, SOM that learns metrics.

4.1 Preprocessing

The raw eye movement data was mapped to fixations using software from the eye movement measuring equipment manufacturer (SMI). The preprocessing is not perfect: since the software is mainly aimed at finding fixations, the saccade measures computed from the fixation data were somewhat inaccurate (for example, blinks will be classified as saccades). The software also smoothes the data temporally and spatially, which means that very short saccades cannot be detected.

The most severe source of errors in the measurements is inaccurate calibration. In order to minimize its effects, eye movement data of each scene was matched to the display by manually finding an affine transformation. In future experiments the matching will naturally be automated. Automatic adjustment of calibration (possibly during a measurement) will be a challenging research problem in itself.

Fixations were mapped to the nearest word. To discard outliers, a fixation was not assigned to any word if the distance to the closest word was 1.5 times the length of the longest word.

Features were computed for each individual word on the display. In order to reduce noise in the experiments, features computed for the words in each title were averaged to a title-specific feature vector. The relevance of each title was specified manually. In order to facilitate the task of finding relevance, the sets where the subject answered incorrectly or where no title containined the answer were left out from the initial data analysis. Before analysing the data, each feature was standardized to have zero mean and unit variance. Due to small amount of data (379 in total; only 41 were associated with titles containing an answer), the number of features was reduced to 21 by leaving out highly correlated features.

4.2 Exploration

We began our analysis with unsupervised methods in Euclidean metrics in order to find out whether there were any 'natural' clusters in the data. We first applied principal component analysis (PCA) to each test subject separately. Since the eigenvectors of all test subjects looked quite similar, we proceeded by combining the data. The first look at the data was then taken by projecting it onto the two principal axes (Figure 1).

Judging from the PCA plot, there seem to be two clusters: One that contains mainly non-relevant titles and another one that is a mixture. Most titles containing the answer belong to the latter cluster together with the other relevant titles. Hence separating them is likely to be a more difficult task. Different users are not separated into clusters, indicating that the used set of standard features is not severely affected by different reading behaviors and strategies.

In order to find out whether there were non-linear

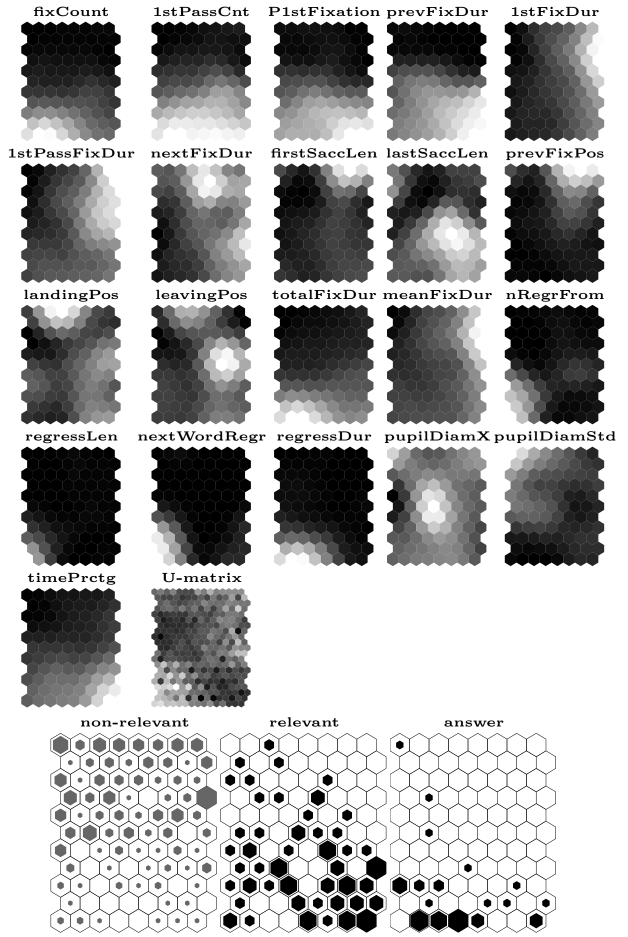

Figure 2: Above: Plot of SOM component planes (see the Appendix). Below: Hits on the SOM of non-relevant and relevant titles, and titles containing the answer. The size of the shadowed hexagon shows the number of data points.

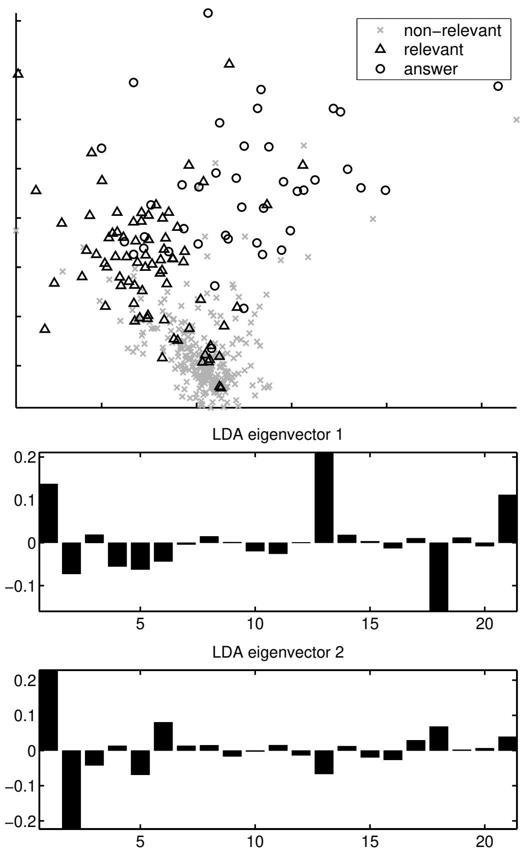

Figure 3: Above: Projection of eye movement data to two discriminative directions (LDA). Below: The normalized eigenvectors of the two discriminative directions. Each bar shows the value of one variable (in the order given in the Appendix) in the eigenvector.

effects in the data, and to evaluate the contribution of each component to clusters, data exploration with SOM was then carried out. Plots of the component planes and the U-matrix are shown in Figure 2. The two clusters which were visible in the PCA plot can be seen also in the U-matrix. Titles of the different degrees of relevance are fairly well separated on the SOM, judging from the classification accuracy that was 82.6 % for the training data. From the component plane plots, it can be seen that the features associated with titles containing the answer are 'fixation count', 'total fixation duration', and 'regression duration', whereas the features associated with relevant titles are 'probability of fixation during first pass', 'previous fixation duration', and 'time percentage'.

4.3 Discriminative components

The clusters found by unsupervised methods already showed an ability to discriminate relevance. Linear discriminant analysis (LDA) was carried out next, since besides (linear) classification, it can also be used for visualizing the data and for evaluating which features are the most important in discriminating the classes. A plot of the data, a projection on the plane defined by the (non-orthogonal) eigenvectors of the best separating directions is shown in Figure 3. The eigenvectors have been plotted as well, indicating that less than ten features seem to be important in the LDA.

From the LDA visualization in Figure 3, it seems that the three classes can be separated. However, since the assumptions of LDA (normal distribution with equal covariances in all classes) are heavily violated in this data, the next step was to test whether the results did truly show an effect by carrying out classification with leave-out data (results are shown in Table 1).

4.4 Supervised exploration

Linear discriminant analysis showed that relevance can be determined to a certain extent with our set of standard features. However, the classification accuracy is not perfect and as the results of SOM already show, there are also non-linear effects in the data. Hence a non-linear method, a SOM that learns metrics, was applied to learn more about which kinds of features are

Table 1: LDA classification to titles containing answer, relevant titles, and non-relevant titles. Results were obtained using leave-one-out validation assuming equal priors for different classes. Boldface: statistically significant difference against a dumb classifier (McNemar's test).

Subject	Accuracy %	(p-val)
Subject 1	**79.7**	(0.018)
Subject 2	68.6	
Subject 3	**77.8**	(0.024)
All subjects	**80.5**	(< 0.001)

useful.

The metric is constructed using auxiliary data, in our case the associated relevance. We first form a conditional probability density estimator for $p(c|\mathbf{x})$, where $\mathbf{x}$ is the original feature vector and c the associated relevance. A local metric matrix is then formed of the Fisher information matrix

$$\mathbf{J}(\mathbf{x}) = -E_{p(c|\mathbf{x})}\left[\frac{\partial^2}{\partial \mathbf{x}^2}\log p(c|\mathbf{x})\right], \qquad (1)$$

with which we can evaluate distances between $\mathbf{x}$ and $\mathbf{x}+d\mathbf{x}$ by $d\mathbf{x}^T\mathbf{J}(\mathbf{x})d\mathbf{x}$. SOM learning is then carried out in this metric (for a detailed description of the method, see [6, 9]).

The results are visualized in Figure 4. Both relevant and answer titles form a focused cluster on the SOM, and non-relevant titles have been scattered onto the top region. The displays of relative importances (the contribution to the metric) of each feature show where the feature is an important factor in explaining class changes. Many of the displays are non-linear, indicating non-linear effects in the data. Many features contribute to separating the relevant titles from the various kinds of non-relevant titles, and some (e.g. 'mean fixation duration' and 'regression duration') to separating the answer titles from the relevant ones. A detailed analysis of different (sub)clusters of data and contributions of features in them will be left for further publications. The selection and implementation of a non-linear classifier will also be studied in the future.

Interestingly, the pupil size measures seem to be relevant for relevance, an effect which was not visible on the ordinary SOM. Results backing up this finding have been reported in [5] where other factors affecting the pupil size, such as luminosity, were more strictly controlled (this level of control is not realizable in practical applications such as proactive information retrieval). Different clusters were more clearly separated in the learned rather than the Euclidean metric; classification accuracy for the training data was now 86.8 %.

Finally, previously discarded data from the sets where the subject answered incorrectly or where there was no answer, was mapped to the SOM that has learned metrics. Number of data samples in each SOM unit, "hits", are shown in Figure 5. The display hints at that the relevant items may get hits closer to the answer area; the reason is possibly that more processing was required from the subjects to ascertain that there really was no answer. The titles containing correct answers became mapped to the non-relevant region, which is in agreement with the users' own decision. Analogously, the titles incorrectly classified by the user as containing the answer became mapped to the answer area. In conclusion, the system monitors the attention of the user, even when she makes mistakes.

5 Discussion

We have introduced an experimental setup for finding relevance from eye movements during a browsing task. Several standard measures used for analysing reading behavior were taken as features and then data analysis was carried out to learn whether it is possible to determine relevance from eye movements, and which features are most important for doing this. In the future we intend to repeat the experiment with more subjects and more assignments.

In this feasibility study we verified that the setup and the standard features are sufficient for determining relevance to some extent for titles (sentences). The ability to estimate relevance for individual words would be desirable as well. For this task, new features need to be introduced, possibly taking into account the correlation of fixation time with word occurrence frequency.

Acknowledgements

The experimental setup was devised by the authors together with Janne Sinkkonen and Kai Puolamäki. We would also like to thank Nelli Salminen for assistance in implementation of the experiments.

This work was supported by the Academy of Finland, grant 200836.

References

[1] J. Beatty and B. Lucero-Wagoner. The pupillary system. In J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, editors, *Handbook of Psychophysiology*, chapter 6. Cambridge University Press, Cambridge, UK, 2000.

[2] M. G. Calvo and E. Meseguer. Eye movements and processing stages in reading: Relative contribution of visual, lexical and contextual factors. *The Spanish Journal of Psychology*, 5:66–77, 2002.

[3] R. Engbert, A. Longtin, and R. Kliegl. A dynamical model of saccde generation in reading based on spatially distributed lexical processing. *Vision Research*, 42:621–636, 2002.

[4] A. Hyrskykari, P. Majaranta, A. Aaltonen, and K.-J. Räihä. Design issues of iDict: a gaze-assisted translation aid. In

Proceedings of Eye Tracking Research and Applications (ETRA2000), pages 9–14. ACM press, 2000.

[5] M. A. Just and P. A. Carpenter. The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, 47:310–339, 1993.

[6] S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

[7] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 2001. Third edition.

[8] S. P. Liversedge and J. M. Findlay. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4:6–14, 2000.

[9] J. Peltonen, A. Klami, and S. Kaski. Learning more accurate metrics for self-organizing maps. In J. R. Dorronsoro, editor, *Artificial Neural Networks—ICANN 2002*, pages 999–1004. Springer, Berlin Heidelberg, 2002.

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

[11] E. D. Reichle, A. Pollatsek, D. L. Fisher, and K. Rayner. Toward a model of eye movement control in reading. *Psychological Review*, 105:125–157, 1998.

[12] D. D. Salvucci. *Mapping eye movements to cognitive processes*. PhD thesis, Department of Computer Science, Carnegie Mellon University, 1999.

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

Appendix

The eye movement features used in this paper:

fixCount: Total number of fixations to the word; **FirstPassCnt**: Number of fixations to the word when the word is first encountered; **P1stFixation**: Did a fixation to a word occur when the sentence that the word was in was encountered for the first time; **prevFixLen**: Duration of the previous fixation when the word is first encountered; **firstFixDur**: Duration of the first fixation when the word is first encountered;. **firstPassFixDur**: Sum of durations of fixations to a word when it is first encountered; **nextFixDur**: Duration of the next fixation when the gaze initially moves on from the word; **firstSaccLen**: Distance (in pixels) between the first fixation on the word and the previous fixation; **lastSaccLen**: Distance (in pixels) between the last fixation on the word and the next fixation; **prevFixPos**: Distance between the fixation preceding the first fixation on a word and the beginning of the word (in pixels); **landingPosition**: Distance of the first fixation on the word from the beginning of the word (in pixels); **leavingPosition**: Distance between the last fixation before leaving the word and the beginning of the word (in pixels); **totalFixDur**: Sum of all durations of fixations to the word; **meanFixDur**: Mean duration of the fixations to the word; **nRegressionsFrom**: Number of regressions leaving from the word; **regressLen**: Sum of durations of fixations during regressions initiating from the word; **nextWordRegress**: Did a regression initiate from the following word; **regressDur**: Sum of the durations of the fixations on the word during a regression; **pupilDiamX**: Mean horizontal pupil diameter during fixations on the word minus mean pupil diameter of the subject during the experiment; **pupilDiamStd**: Standard deviation of the pupil horizontal diameter during fixations on the word; **timePrctg**: First fixation duration divided by the total duration of fixations on the display.

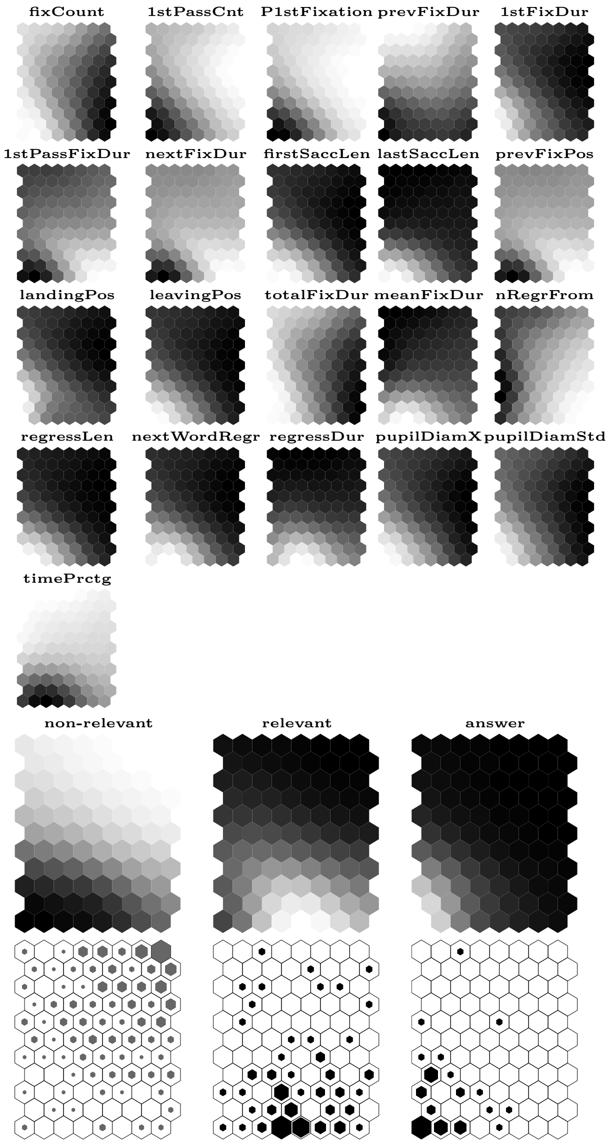

Figure 4: Above: Visualization of the relative contribution of each feature to the approximative Fisher metric at each SOM unit. Center: A plot of conditional probability densities in each node for non-relevant and relevant titles, and titles containing the answer. Below: Number of data points in each SOM unit.

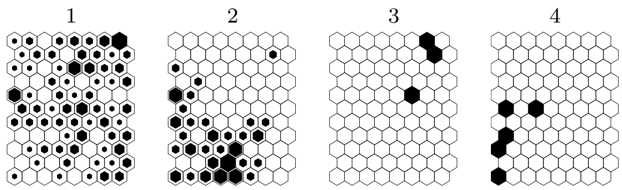

Figure 5: Hits on SOM with learning metrics for data where there was no answer or where the subject's answer was incorrect. 1: non-relevant titles, 2: relevant titles, 3: titles containing an answer, 4: incorrect answers of the subjects.