Evaluating query-independent object features for relevancy prediction*

Andres R. Masegosa[†], Hideo Joho[‡], and Joemon M. Jose[‡]

andrew@decsai.ugr.es, {hideo,jj}@dcs.gla.ac.uk
[†]Department of Computer Science and A.I., University of Granada, Spain.
[‡]Department of Computing Science, University of Glasgow, UK.

Abstract. This paper presents a series of experiments investigating the effectiveness of query-independent features extracted from retrieved objects to predict relevancy. Features were grouped into a set of conceptual categories, and individually evaluated based on click-through data collected in a laboratory-setting user study. The results showed that while textual and visual features were useful for relevancy prediction in a topic-independent condition, a range of features can be effective when topic knowledge was available. We also re-visited the original study from the perspective of significant features identified by our experiments.

1 Introduction

There has been a growing interest in leveraging *contexts* in different aspects of Interactive Information Retrieval (IIR) systems [1–3]. While the IR community might not have a consensus regarding what exactly a context is, the progress has been made on the understanding of IR in contexts. For example, Ingwersen and Järvelin [4] propose a model of context stratification which includes a wide range of features in the information seeking and retrieval environment. The model offers structured focus for the work on finding the potentially significant contexts to improve the performance of IIR systems. Some of the proposed strata relevant to this work are: work task features; interaction features; and document features.

One way to identify significant contextual features is to investigate their relationship to the relevancy of retrieved objects. For example, Kelly and Belkin [5] found that the reading time of documents can vary significantly across the topics, thus, it can be difficult to predict the document relevancy. Fox et al. [6] applied a machine learning technique to model the interaction features with respect to the document relevancy. Another way to find significant features is to observe the effect of features in an IR technique such as relevance feedback. For instance, White, et al. [7] investigated the effects of topic complexity, search experience, and search stage in the performance of implicit relevance feedback. Furthermore, the relationship between the context strata is important to understand the significance of features. For example, Freund, et al. [8] suggest that the document genres can be indicative of the type of topics in a workplace environment.

* This work was supported by ALGRA project (TIN2004-06204-C03-02), FPU scholarship (AP2004-4678) and EPSRC (Ref: EP/C004108/1)

Table 1. Conceptual categories for object features.

Main Category	Sub category	Code	Feature size
Document textual features		DOC	14
Visual/Layout features	Visual Appearance	V-VS	28
	Visual HTML tags	V-TG	27
	Visual HTML attributes	V-AT	16
Structural features		STR	18
Other selective features	Selective words in anchor texts	O-AC	11
	Selective words in document	O-WD	11
	Selective HTML tags	O-TG	7
	Selective HTML attributes	O-AT	16

In this paper, we present a series of experiments investigating the effectiveness of query-independent object features to predict document relevancy. Our evaluation was based on experimental data collected in a laboratory-based user study with 24 participants searching four different topics. Participants were asked to bookmark a document when perceived relevant information was found. Object features were extracted from click-through documents and bookmarked documents. The objective of using such data was two-fold: First, we can decrease the uncertainty of users' underlying information needs inferred from interaction data, compared to a search engine's transaction log. Second, we were interested in re-visiting the result of the original study from the perspective of significant features identified by the experiments.

The rest of this paper is structured as follows. Section 2 describes our methodology to extract a set of potential object features, and other operations to improve the performance of the relevancy prediction. Section 3 presents a series of experiments investigating the effectiveness of the object features. Section 4 discusses the implications of our findings. Finally, Section 5 concludes the paper with future work.

2 Methodology

This section describes our methodology used to extract object features and other operations applied to the features to improve the performance of relevancy prediction. An overview of the classifiers used in our experiment and performance measures is also discussed.

2.1 Conceptual categories of object features

The first step to find effective features for relevancy prediction was to identify candidate features that can be extracted from retrieved documents. Based on some informal experimentation and literature survey, we have identified approximately 150 object features. To increase the understanding of candidate features in relevancy prediction, we then grouped them into a set of conceptual feature categories. The structure used for the categorisation is shown in Table 1.

As can be seen, there are four main categories: Document textual features, visual/layout features, structural features, and other selective features. The objective of the main categories was to group candidate features into a set of independent functionality played in a document. Therefore, we do not claim that our categorisation is ideal for all applications. On the contrary, the structure of features should be revised as an investigation progresses. Nevertheless, the structure shown in Table 1 was a good starting point for us to investigate the effectiveness of object features. An overview of the main categories is as follows.

Document textual features: This category consists of features that are related to textual contents of documents. The examples of features include the number of words in a document and anchor texts, number of upper-case words, number of digits, Shannon's entropy value [9] for a document and anchor texts.

Visual/Layout features: This category consists of features that are related to visual or layout aspects of documents. There are three sub categories: *Visual appearance* includes features such as the number and dimension of (background) images and foreground/background colours; *HTML Tags* includes a set of HTML tags such as `font`, `li`, and `table`; *HTML attributes* includes attributes used across the HTML tags such as `style`, `border`, and `face`.

Structural features: This category consists of features that are related to hyperlink and site structure of documents. The examples include the depth of document in a URL, the number of outlinks, PageRank scores.

Other selective features: This category consists of features that are not necessarily fit into the above categories. There are four sub categories: *Selective words in document* includes the presence of selective words such as address, search, and help; Selective words in anchor texts is the same as above but extracted from only anchor texts; Selective HTML tags includes a set of HTML tags such as `form`, `object`, and `script`; Selective HTML attributes includes `lang`, `onclick`, `src`, etc.

Extraction of the features were carried out by a mixture of tools such as [10] and [11]. The following sections describe a proposed methodology to build classifier, to select significant features, and finally, to validate the results.

2.2 Probabilistic Classification Approach

The classification problem can be seen as an ability of predicting a given feature of an object using another set of features of the same object. In the probabilistic classification paradigm, the classification problem can be described by two types of random variables:

Class Variable: C. This random variable is the variable to be predicted. This variable contains one state for each possible prediction $\Omega_C = \{c_1, ..., c_k\}$. In our case, $\Omega_C = \{Relevance, Non - Relevance\}$.

Predictive/Attribute Variables: $\mathbf{X} = \{X_1, X_2, ..., X_n\}$. Each variable has a set of possible states (discrete variables) or continuous values (continuous variables). For simplicity, we are only going to talk about discrete variables. Then, $\Omega_{X_i} = \{x_{ij_1}, ..., x_{ij_k}\}$ is the set of possible states of the X_i random variable. In our case, $\mathbf{X}$ is the set of variables described in Section 2.1.

In this model, our objective is to learn the following probability distribution:

$$\mathbf{P}(C|X_1, X_2, ..., X_n) = \{\mathbf{P}(c_1|X_1, ..., X_n), ..., \mathbf{P}(c_k|X_1, ..., X_n)\}$$

In other words, the probability of the class variable given the set of attributes variables. The prediction of the class (Relevance or Non-Relevance) is based on the highest *a posteriori* probability. This probability distribution is estimated by a set of data $\mathbf{D} = \{D_1, ..., D_M\}$, where each D_i contains an instantiation of the predictive features and the class for the object number i:

$$D_i = \{x_{1j_1}, x_{2j_2}, ..., x_{nj_n}, c_j\}$$

In our study, a great number of attribute variables were continuous. However, the literature suggests that the performance of the classifiers can be more robust when the variables are discrete data. Therefore, we used the *equal frequency* discretisation method [12] to split the continuous variables into 10 intervals.

Another aspect to consider in the classification was the balance of the class variable distribution (i.e., the portion of relevant and non-relevant documents in a data set). An imbalance data is known to deteriorate the performance of a classifier [13]. We took the following approach to address the issue. When there were a large number of cases, we randomly removed the cases from the larger class until the portion was balanced. When there were a small number of cases, we used a resampling method to balance the data. Although this resampling method was a good technique to correct imbalanced data, it was also possible to over-estimate the performance of the classifier. We used AUC measure [12] to detect the over-estimation.

2.3 Classifiers Used

While a single Bayesian network approach was used by [6], we were interested in using several classifiers and reporting the result of the best performing classifier. This was because a single classifier was unlikely to show the significance of attribute variables in a complex dependency structure. We selected four classifiers that have been proved to be successful in machine learning classification. An overview of the classifiers used in our experiments is as follows.

Naive Bayes [14] This is one of the well known probabilistic classifiers. The main assumption in this model is that all attribute variables depend on the class variable and they are independent of each other.

AODE [15] This classifier can use multiple representations of a problem space to predict the class variable. A disadvantage is that this classifier can not show an explicit relationship between variables.

HNB [16] This classifier creates a hypothetical variable to represent the relationship between the attribute variables. The resulted representation is then used to predict the class variable. HNB inherits the structural simplicity of Naive Bayes and can be trained without mining the dependency structure.

K2-MDL This classifier is a variant of Bayesian networks classifier [17] where the structure is learnt by the K2 algorithm [18].

2.4 Feature selection scheme

The feature selection in the supervised classification paradigm is to find a minimum set of attribute variables that can achieve the best performance. The selection of significant features in the problem space can prevent the classifiers from introducing noisy evidences in the training stage. The feature selection can also reduce the number of variables to be considered in the problem space, thus, it can facilitate our understanding of significant variables.

While several techniques have been proposed for the feature selection [19], we used a wrapper method which can select a set of the best features based on the AODE classifier. The actual selection process was similar to the cross validation method described in the following section. The final set of features was generated by the features that were selected at least N% of the repeated cross validation process. We used 50%, 80%, and 90% as the cutoff levels in the feature selection. We found that the overall performance did not vary significantly over the cutoff levels. Therefore, we only report the results of 90% in the experiment since it consists of the smallest number of significant features.

2.5 Classification Validation Scheme

With the aim of provide a robust estimation of the accuracy of a classifier, the set of data was partitioned in two separated sets. The training data set was used to build the classifier and the test data set was used to estimate the performance. The K-fold-cross validation method was used to partition the data set as follows. The data set $\mathbf{D}$ was divided in K random subsets with the same size $\{D_1, ..., D_K\}$, thus, the validation process was repeated K times. In other words, in the step i with $i = 1...K$ a training data set was defined $T_i = \mathbf{D} \setminus D_i$ and the subset D_i was used as a test set and the accuracy was measured based on them. The mean of the K accuracy measures was reported as the final estimated performance of the classifier. In our study, a 10 fold-cross validation was repeated 10 times to measure the performance (i.e., based on 100 repeated estimations).

3 Experiments

This section presents a series of experiments which investigated the effectiveness of query-independent contextual features to predict the relevancy of click-through documents. The accuracy of prediction is defined by the portion of correct prediction in the total number of click-through documents. The correct prediction is a sum of the true positive and true negative cases (i.e., predicting a relevant document as relevant, and predicting a non-relevant as non-relevant). For example, when the data consist of 50 relevant and 50 non-relevant documents, and when 30 relevant and 40 non-relevant documents are correctly predicted, then the performance is 70%[1].

Throughout this section, the results are presented in two groups of data set. The first group is based on all click-through data without the distinction of individual topics. The second group is based on the data within individual topics. The former is referred to as the *topic-independent* set, and the latter is referred to as the *topic-dependent* set. This

[1] $\frac{30R+40NR}{50R+50NR} = \frac{70}{100} = .7$

Table 2. Baseline performance of relevancy prediction.

	Click-through	Relevant	Non-Rel	Baseline (%)	Balanced (%)
No topic	737	375	362	50.9	
Topic 1	203	123	80	60.6	50.0
Topic 2	173	83	90	52.1	
Topic 3	154	69	85	55.2	
Topic 4	207	100	107	51.7	

grouping enables us to examine the effect of topic knowledge in relevancy prediction, and how the effectiveness of the features differs in the two conditions.

The section is structured as follows. First, the baseline performance of relevancy prediction is established by looking at the portion of relevant/non-relevant documents in the click-through data set. Second, the effect of contextual features in each category is examined. Then, the effects of several operations on the contextual features are presented: feature selection, feature combination, and use of highly relevant documents.

3.1 Baseline performance

A total of 1038 click-through documents were extracted from our user study of 24 participants searching four different search topics [20]. Of those, 375 were unique relevant and 362 were unique non-relevant documents. Therefore, the baseline performance of relevancy prediction was set to 50.9% in the topic-independent set (denoted as *No topic* in the tables). The portion of relevant/non-relevant documents varied across the four topics. The baseline performance was taken from whichever the higher portion of relevance, as shown in Table 2.

Note that a relatively large difference was found between the number of relevant and non-relevant documents in Topic 1. To measure an accurate performance of the classifiers, we generated a balanced data set by a random sampling for Topic 1 (shown in the 6th column of Table 2). In the following analysis, the performance based on the balanced set is used for Topic 1. No change was found to be necessarily for the rest of the data set.

3.2 Effect of contextual features

The first experiment examined the effect of contextual features in the individual context categories. In this experiment, the classifiers used only the features defined in each category to predict the relevancy, and the same procedure was repeated for all categories. The performance of relevancy prediction was compared to the baseline performance and the relative improvement was shown in Table 3. The bottom row of the table shows the average improvement across the four topics (but not including the topic-independent set). The statistically significant differences are highlighted in bold in the table. We used the t-test ($p \leq .05$) for the statistical tests throughout the study.

As can be seen, the features in the DOC and V-AT categories were found to be useful for improving the relevancy prediction in the topic-independent set. While the V-VS

Table 3. Effect of contextual features.

		DOC	V-VS	V-TG	V-AT	STR	O-AC	O-WD	O-TG	O-AT	Mean
No topic	50.9	**+5.7**	+2.5	+2.2	**+3.9**	+0.4	+1.2	-2.1	-0.9	+2.2	+1.7
Topic 1	50.0	+4.1	**+11.3**	-2.7	+4.2	+2.0	+5.7	-4.0	+3.1	+2.7	+2.9
Topic 2	52.1	+0.3	+7.3	+6.1	-2.2	+2.9	-9.2	-4.3	+2.0	+4.8	+0.9
Topic 3	55.2	+0.6	+4.4	-2.6	+5.4	+1.1	+8.1	-0.8	+4.8	+6.3	+3.0
Topic 4	51.7	-5.1	-2.6	+0.6	+2.3	+2.1	+1.4	+3.1	+0.6	-0.7	+0.2
Mean		+1.1	+4.6	+0.7	+2.7	+1.7	+1.4	-1.6	+1.9	+3.1	+1.7

Table 4. Effect of feature selection with 90% cutoff.

		DOC	V-VS	V-TG	V-AT	STR	O-AC	O-WD	O-TG	O-AT	Mean
No topic	50.9	**+4.6**	+1.4	+3.1	**+3.8**	+2.4	+2.9	+2.0	+0.3	+1.1	+2.4
Topic 1	50.0	0.0	+1.4	**+10.8**	+7.4	**+11.5**	+6.1	+5.8	+4.1	+3.8	+5.6
Topic 2	52.1	0.0	-3.8	+5.8	+4.8	+5.2	+3.7	+0.9	+2.6	-9.0	+1.1
Topic 3	55.2	0.0	**+8.4**	0.0	+6.2	+0.7	**+10.1**	-1.4	+6.4	-3.3	+3.0
Topic 4	51.2	+2.3	0.0	+3.5	+1.3	+5.8	+5.4	**+8.6**	-1.5	+5.4	+3.4
Mean		+1.4	+1.5	+4.6	+4.7	+5.1	+5.6	+3.2	+2.4	-0.4	+3.1

category was found to be effective in Topic 1, the overall effect of individual categories appeared to be weak across the topics. Furthermore, the performance of most categories appeared to be inconsistent across the topics. The exceptions were the STR and O-TG categories, but the differences were not significant. The following sections present the effects of several operations on the features to improve the performance.

3.3 Effect of feature selection

In the previous experiment, all features were used to predict the document relevancy in the individual categories. One way to improve the performance is to use only a subset of features that are likely to contribute to the prediction, which is called a *feature selection*. There are several methods of the feature selection. In this study, we used the features that were selected 90% of times in the repeated tests on the training set. The feature selection was carried out in the indivdual categories and the result of relevancy prediction is shown in Table 4.

From the far right column (Mean) of Table 4, there appears to be an overall positive effect of the feature selection, compared to Table 3. However, in the topic-independent set, the performance of the significant categories (i.e., DOC and V-AT) was degraded by the feature selection. This suggests that a greater number of the features should be considered in the individual categories when no topic knowledge was available for the relevancy prediction.

On the other hand, a significant improvement was found in several categories of the topic-dependent sets when the feature selection was carried out. The results show that, for example, Topic 1 is likely to benefit from the features in the V-TG and STR categories. What is more important is that the significant category is likely differ across the topics. In fact, no single category contributed to a significant improvement on more

Table 5. Effect of feature combination.

		Best Cat	Combined	High Rel
No topic	50.9	**+5.7**	**+5.6**	**+11.6**
Topic 1	50.0	**+11.5**	**+10.3**	
Topic 2	52.1	+5.8	**+9.2**	
Topic 3	55.2	**+10.1**	**+9.9**	
Topic 4	51.7	**+8.6**	+2.2	
Mean			+7.4	

than one topic. Topic 2 appeared to be a particularly difficult topic to find effective features. This suggests that the effectiveness of query-independent features is fairly topic-dependent.

3.4 Effect of feature combination

So far, we have examined the performance of the individual feature categories. The features selected in the previous experiment appeared to have a varied effectiveness over the topics. We further investigated the performance of the context features by combining the selected features into a single category. The advantage of the feature combination is that the classifiers do not have to find a particular category for the relevancy prediction. The result is shown in Table 5. In the table, the performance of the best category in the previous experiment is shown in the 3rd column (Best Cat), and the performance of the combined feature is shown in the 4th column.

As can be seen, the overall performance of the feature combination appears to be comparable to the best performing category in the previous experiment, except Topic 4. And, the effect appears to be consistent across the two data sets. In particular, Topic 2 was found to benefit from the feature combination significantly. The mean value of four topics (the bottom row of Table 5) suggests that the performance of combined feature is likely to be more consistent than any single feature category. We also tested the different cutoff levels (50%, 80%, and 90%) of the feature selection before the combination, and a similar performance was found over the cutoff levels.

3.5 Effect of highly relevant documents

The last experiment in this paper looked at the effect of highly relevant documents for the relevancy prediction. In the literature, the importance of highly relevant documents has been suggested in the evaluation of IR systems [21]. In this study, the highly relevant documents were determined when the document was judged as relevant by at least two participants in the same topic. While this criterion was not based on a graded relevance judgement, it enabled us to select a reasonable number of relevant documents whose relevancy was shared by participants.

There were a total of 96 documents that were judged by at least two participants. Of those, 69 were relevant and 27 were non-relevant. Similar to Topic 1's data set, we needed to balance the portion of relevant/non-relevant documents for the this analysis.

Table 6. Effect of highly relevant documents (without feature selection).

		DOC	V-VS	V-TG	V-AT	STR	O-AC	O-WD	O-TG	O-AT	Mean
All rel	50.9	**+5.7**	+2.5	+2.2	**+3.9**	+0.4	+1.2	-2.1	-0.9	+2.2	+1.7
High rel	50.0	**+15.2**	**+15.5**	**+16.7**	+8.2	†+12.0	+4.9	+0.3	**+11.6**	+6.2	+9.8

†An over-estimation was detected in this result, thus, not considered.

We used a replacement method which was often used in machine learning (See Section 2.2). The method is known to be robust to measure the performance in a similar situation, and to detect any anomalies in the results.

The effect of highly relevant documents is shown in Table 6. We only report the result of the topic-independent set since the data was too small to measure the individual topic performance. As can be seen, the significant improvements were found in several categories when the highly relevant documents were targeted for the relevancy prediction. The result also shows that the features from a wider range of categories can be considered for the prediction in the topic-independent set. We also measured the effect of feature combination based on the highly relevant documents, and the result is shown in the 5th column of Table 5. As can be seen, a respectable improvement can be achieved without selecting the best performing category. These results show that the use of highly relevant documents can be a more effective way to predict the document relevancy than the other methods examined in this study. This is interesting because the classifiers usually perform worse when the size of the training data decreases.

4 Discussion

This section discusses the implications of our experimental results. We also re-visit the result of the original user study from the perspective of the significant features identified by the experiments.

4.1 Effectiveness of query-independent features

The findings from our experiments have several implications for the use of query-independent object features to predict the document relevancy. First, the set of effective features can be different when the prediction is performed with/without topic knowledge. In the topic-independent set, the textual document features and visual/layout HTML attributes are likely to be significant to predict the document relevancy. In our experimental conditions, the feature selection or feature combination were found to make little improvement on the performance. However, a simple filtering to select highly relevant documents was found to be effective to improve the performance. In the topic-dependent set, on the other hand, many categories can be effective for the relevancy prediction. However, the effectiveness of individual categories can vary across the topics. The results show that the feature selection and feature combination can be effective for improving the performance in this set. Another implication of our results is that an additional classifier for a topic detection should be used supplementary to the relevance prediction. Such a two-stage approach would allow us to use a topic-dependent

Table 7. Minimum set of significant features[†]

	Topic 1	Topic 2	Topic 3	Topic 4
Topic	Dust allergy in workplace	Music piracy on Internet	Petrol price change	Art galleries and museums in Rome
DOC				
V-VS			**imageBDiskSize**	
V-TG	**meta**, **li**			
V-AT				
STR	**URL-Levels** **HtmlLink**	PR-Page, link URL-Domain numlinksAnchor		
O-AC	contact email search		**contact** **help** **email**	search help tel
O-WD				**search**, **address**
O-TG				
O-AT				

[†]The significant features with 90% cutoff is highlighted with bold. The rest are based on 80% cutoff in the feature selection.

significant category effectively, thus, can be promising to improve the performance. Our preliminary test to predict the four topics using the relevancy classifier showed an accuracy of between 45 to 60% with the average of 55%. A further investigation is under way for the integrated approach.

The results also suggest that the textual document features such as the entropy are rarely effective within the individual topics. This was contrast to their performance in the topic-independent set. Our speculation is that the entropy and other document level features might have a low discriminating power to separate relevant documents from non-relevant. Other features that occur less frequently in the data set appear to have a higher discriminating power. Therefore, a similar phenomenon that motivates the idea of inverse document frequency [22] might be applicable to indicate the significance of query-independent features. This also supports our approach to use a range of objects' features to predict the document relevancy.

4.2 Re-examination of the original study

Our experiments were based on the experimental results of a user study carried out in a laboratory setting. A motivation for using such data was to decrease an uncertainty of users' underlying information needs in the experiments. A distinct objective was to re-examine the result of the original study from the perspective of the significant features identified by the experiments. In this section, we discuss the findings of such analysis. Table 7 shows the minimum set of the query-independent features that contributed to a significant improvement in the individual topics. The minimum set was determined by the multiple cutoff levels (80% and 90%, See Section 2.4) in the feature selection to increase the number of indicative features.

In Topic 1, participants were asked to find the information on the potential solution to dust allergy in a workplace. Some perceived relevant documents contained a list of steps to reduce the dust inside a building. Therefore, the `li` tag in the visual feature category was a significant indicator of the document relevancy. The depth of document in a web site appeared to vary in this topic compared to the other topics. As discussed before, Topic 2 was a difficult topic to find significant features. In this topic, participants were asked to find the information on the damage of music piracy on Internet. The structural features in this topic suggest that participants were able to find relevant information in the top ranked documents from a limited number of URL domains. The most unexpected result was Topic 3 where the disk size of background images (imageBDiskSize) was found to be a significant indicator of the document relevancy. A close examination showed that the background image information was more helpful for predicting non-relevant documents than relevant documents. The selective words in the anchor texts appeared to be useful for this topic. The result of Topic 4 was also interesting. We initially expected that the visual features were likely to be significant in this topic, but this was not the case. Instead, the selective words in a document were found to be a significant indicator of the document relevancy. We speculate that since most click-through documents contained a variety of images in this topic, their discriminating power was lower than we had expected. However, since participants were asked to find the information on a particular location, the words such as *address* was found to be significant.

As can be seen, the result of the re-examination of the original work was a mixture of re-assurance and puzzlement. More importantly, however, the significant features appeared to offer us a pointer for the further examination of the original study. In this sense, the re-examination of the original work based on the significant features can supplementarily used in the evaluation of user studies.

5 Conclusion and future work

This paper presented a series of experiments which investigated the effectiveness of the query-independent features to predict the document relevancy. The experimental results from a user study were used to extract the various features of retrieved objects. Our results show that the document-level textual features and visual features can be indicative of the document relevancy in an topic-independent situation. The use of highly relevant documents can improve the performance significantly. When the type of topics was known, a wider range of features can be effective for the relevancy prediction. However, the effectiveness of the features is likely to vary across the topics. Overall, these findings highlight the importance of investigating the significance of objects' features from the perspective of the topics and aggregated relevance assessments.

In this study, we investigated the features from retrieved objects. We are conducting a similar experiment based on the interaction features in the other parts of the user logs, searchers features gained from the participants" background information and finally, subjective perceptions on the topic characteristics established by the questionnaires. We anticipate that the features from the additional context strata can facilitate the understanding of the original user study. We also plan to evaluate the features extracted

from another user study to investigate the robustness of the significant features identified by this study.

References

1. Ingwersen, P., Belkin, N.: Information retrieval in context - IRiX: workshop at SIGIR 2004. SIGIR Forum **38**(2) (2004) 50–52
2. Ingwersen, P., Järvelin, K.: Information retrieval in context: IRiX. SIGIR Forum **39**(2) (2005) 31–39
3. Ruthven, I., Borlund, P., Ingwersen, P., Belkin, N., Tombros, A., Vakkari, P., eds.: Proceedings of the 1st IIiX Symposium, Copenhagen, Denmark (2006)
4. Ingwersen, P., Järvelin, K.: The Turn: Integration of Information Seeking and Retrieval in Context. Springer (2006)
5. Kelly, D., Belkin, N.J.: Display time as implicit feedback: understanding task effects. In: Proceedings of the 27th SIGIR Conference, Sheffield, United Kingdom, ACM Press (2004) 377–384 1009057.
6. Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve web search. ACM Transactions on Information Systems **23**(2) (2005) 147–168 1059982.
7. White, R.W., Ruthven, I., Jose, J.M.: A study of factors affecting the utility of implicit relevance feedback. In: Proceedings of the 28th SIGIR Conference, Salvador, Brazil, ACM (2005) 35–42
8. Freund, L., Toms, E.G., Clarke, C.L.A.: Modeling task-genre relationships for ir in the workspace. In: Proceedings of the 28th SIGIR Conference, Salvador, Brazil, ACM (2005) 441–448
9. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana, Illinois (1949)
10. : (Html parser)
11. : (Firefox add-ons)
12. Duda, R.O., Hart, P.E.: Pattern Classification. Wiley Interscience (2000)
13. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5) (2002) 429–449
14. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. John Wiley Sons, New York (1973)
15. Webb, G.I., Boughton, J.R., Wang, Z.: Not so naive bayes: aggregating one-dependence estimators. Mach. Learn. **58**(1) (2005) 5–24
16. H. Zhang, L.J., Su, J.: Hidden naive bayes. In: Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05), (AAAI Press(2005).)
17. Pearl, J.: Probabilistic Reasoning with Intelligent Systems. Morgan & Kaufman, San Mateo (1988)
18. Cooper, G., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine Learning **9** (1992) 309–347
19. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(1-2) (1997) 273–324
20. Joho, H., Jose, J.M.: Slicing and dicing the information space using local contexts. In: Proceedings of the First Symposium on Information Interaction in Context (IIiX), Copenhagen, Denmark (2006) 111–126
21. Järvelin, K., Kekäläinen, J.: Ir evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd SIGIR Conference., Athens, Greece, ACM (2000) 41–48
22. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28**(1) (1972) 11–21