Received 13 January 2006
Revised 13 June 2006
Accepted 24 June 2006

The relative effects of knowledge, interest and confidence in assessing relevance

Ian Ruthven, Mark Baillie and David Elsweiler
Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

Abstract

Purpose – The purpose of this paper is to examine how different aspects of an assessor's context, in particular their knowledge of a search topic, their interest in the search topic and their confidence in assessing relevance for a topic, affect the relevance judgements made and the assessor's ability to predict which documents they will assess as being relevant.

Design/methodology/approach – The study was conducted as part of the Text REtrieval Conference (TREC) HARD track. Using a specially constructed questionnaire information was sought on TREC assessors' personal context and, using the TREC assessments gathered, the responses were correlated to the questionnaire questions and the final relevance decisions.

Findings – This study found that each of the three factors (interest, knowledge and confidence) had an affect on how many documents were assessed as relevant and the balance between how many documents were marked as marginally or highly relevant. Also these factors are shown to affect an assessors' ability to predict what information they will finally mark as being relevant.

Research limitations/implications – The major limitation is that the research is conducted within the TREC initiative. This means that we can report on results but cannot report on discussions with the assessors. The research implications are numerous but mainly on the effect of personal context on the outcomes of a user study.

Practical implications – One major consequence is that we should take more account of how we construct search tasks for IIR evaluation to create tasks that are interesting and relevant to experimental subjects.

Originality/value – Examining different search variables within one study to compare the relative effects on these variables on the search outcomes.

Keywords Information retrieval, Information searches, Retrieval performance evaluation, Search output, Cognition, Information operations

Paper type Research paper

1. Introduction

Understanding how people assess relevance has been one of the core research areas in information retrieval (IR) since its inception as an academic discipline. If we understand more about the relevance decisions people make when searching then we can construct interactive systems that facilitate making these decisions, or systems that make better predictions regarding user search behaviour. We can also better understand how to evaluate the effectiveness of search systems to individual searchers by understanding what searchers intend by an assessment of relevance.

In this paper we examine how an assessor's interest in a search topic, their knowledge about a search topic and their confidence in assessing relevance can change the way in which the assessor assesses relevance. We do this with specific reference to

Journal of Documentation
Vol. 63 No. 4, 2007
pp. 482-504

0022-0418
DOI 10.1108/00220410710758986

an experimental study we carried out as part of our participation in this year's HARD track of the Text REtrieval Conference (TREC) (Allan, 2005; Voorhees and Buckland, 2006). What our study shows is that an assessor's personal context, that is their knowledge of and attitudes to a search task, affect how they assess relevance. Our results indicate that assessors with high knowledge regarding a search topic, high interest in the search topic or high confidence in assessing documents will assess more documents as being relevant to a search than assessors with lower topical knowledge, interest or confidence. We also present results on how these personal factors affect an assessor's ability to predict what information they might find useful. Finally, we consider which of these factors it is useful to know about an assessor and discuss the implications of our results for the design and evaluation of interactive information retrieval systems

The remainder of the paper is structured as follows: in section 3 we give a short introduction to the HARD track as background to our research, in section 4 we present the details of our study and in section 5 we present our findings. Prior to this, in section 2, we describe previous work on relevance assessment behaviour.

2. Related work

Relevance is the core concept in IR (Borlund, 2003; Mizzaro, 1997; Ruthven, 2005). Researchers have investigated how people assess the relevance of documents either for the purpose of understanding human search behaviour or for the purpose of improving algorithms, such as relevance feedback algorithms, that utilise human assessments of relevance. Assessing relevance would appear to be a simple process: either a document is relevant or not to a searcher's task or information need. However, as Katter (1968, p. 1) noted, very early on, a "recurring finding from studies involving relevance judgments is that the inter- and intra-judge reliability of relevance judgments is not very high". That is, the same person may judge the same document relevant or non-relevant at different points in time, even within a single search session. Different searchers may also disagree on the relevance of a document to a search request.

Relevance would then seem to be a very ephemeral concept upon which to base a research discipline. However, the more we examine how people assess relevance the more we can understand why this inconsistency in judging relevance occurs and how it reflects the subjective and contextual nature of the assessment process.

Inconsistency in relevance judgements can occur for a number of reasons, for example:

- *The measurement of relevance is affected by how we measure relevance*. To investigate relevance, we need to be able to measure it in some way and a standard practice is to ask searchers to estimate the relevance of a document using some substitute representation, e.g. ordinal scales, relevance categories or, most commonly, using a simple binary relevant/not relevant decision (Mizzaro, 1999). How we ask searchers to record the relevance of individual items can affect our understanding of the searchers' thought processes and evaluation measures that use relevance. For example, Eisenberg and Hu (1987) demonstrated that binary relevance decisions, asserting that a document is relevant or not-relevant, can distort measures such as recall and precision because the point at which individual searchers distinguish between relevant and non-relevant is not consistent. That is, searchers may mark the same

document relevant or not relevant for the same search because each searcher has a different threshold for deciding relevance. Allowing searchers to order the retrieved documents is not a solution either because this may force searchers to make a distinction between documents where the searcher sees no difference (Katter, 1968).

Also, measurements of relevance do not take into account that objects may be relevant for different reasons, e.g. Barry and Schamber (1998), Tombros *et al.* (2005) and these reasons may change over time (Choi and Rasmussen, 2002; Vakkari and Hakala, 2000). These reasons, or relevance criteria, are not always easily compared by a searcher, e.g. is the recency of a document more or less important than its author? Ignoring why assessments are made runs the risk of misinterpreting what searchers view as important about retrieved material. Other aspects of the assessment process that could affect the assessments include whether the assessments are measured against a search request or against an underlying information need (Mizzaro, 1997) and the representation of the information objects used, e.g. assessing relevance against whole documents such as full-text articles or against surrogates such as abstracts. Even the mode of delivery can affect judgements of relevance, Tombros and Crestani, for example showed different patterns of assessment on spoken versus textual versions of the same documents (Tombros and Crestani, 2000).

- *Relevance is affected by context of the search.* Measurements of relevance and relevance-based evaluation criteria commonly assume the relevance of one document is independent of the relevance of other retrieved objects (Buckley and Voorhees, 2005). However, the relevance of a document is not normally assessed in isolation but within the context of an interactive search and the assessment of one document is affected by the presence and relevance of other document. Several studies, e.g. Eisenberg and Barry (1998), Florance and Marchionini (1995), and Huang and Wang (2004) have demonstrated that searchers assess the relevance of a document relative to the relevance of documents they have already seen. The assignment of a relevance decision to a document is specifically affected by the number of relevant documents already seen in a search, and viewing one relevant document can change the searcher's perception of the relevance of subsequently viewed documents. This can mean, for example as demonstrated by Vakkari and Sormunen (2004) and Florance and Marchionini (1995), that if a search returns very few highly relevant documents, a searcher may mark marginally relevant documents as being relevant. On the other hand, if a search returns many relevant documents, the searcher may mark marginally relevant documents as not relevant.

  Documents can also possess relationships to each other which counter the notion that documents should be seen as independently relevant, e.g. collections can contain duplicate documents, documents may refer to other documents in the same collection and the searcher may need to understand one document to be able to interpret another. Tiamiyu and Ajiferuke (1988) showed that different interpretations of the inter-relatedness of documents could affect the results of evaluations using standard recall and precision figures. The degree of relevance itself is of importance to evaluating IR systems and several authors have investigated allowing searchers to express the degree of relevance of a document

to a search (Järvelin and Kekäläinen, 2000; Lesk and Salton, 1969; Voorhees, 2001). What they have typically shown, and what we also show in this paper, is that allowing for a relative assessment of relevance gives us a more accurate understanding of the performance of IR systems.

- *Relevance is affected by context of the searcher*. A search is an intellectual task and one that is affected by characteristics of the searchers themselves, their relationship to the search task and to the documents retrieved. In investigating non-binary, or partial, relevance assessments Spink *et al.* (1998) investigated differences in how searchers used judgements of high and partial relevance. Their findings demonstrate the range of reasons why a searcher may assess a document as only partially relevant and that these reasons may be due to the content of the document ("not enough information", p. 611), the relation of the document to other documents ("duplicate information", p. 611) or the searchers' understanding of their own information problems.

Of particular interest to this paper, their analysis indicates that although individual assessments of partial relevance can reflect poor match between an information object and an information problem, high numbers of partial assessments within a search stage can be indicative of changes in the searcher's perception of their own search needs. That is, high numbers of partial relevance assessments can correspond to a stage where assessors are unsure of how they should define relevance within a particular search stage because their understanding of their information problem, and possibly their criteria for judging relevance, had changed as a result of searching. In one of the studies described in Spink *et al.* (1998), searchers' use of partial assessments were not a confident assessment of the relevance of the information. Rather they signified a stage where the searcher was delaying making a confident decision of relevance until they had a clearer idea of the task for which the relevant information would be used ("Could be helpful but I don't know yet", p. 611).

In a separate, and longitudinal, investigation Vakkari and Hakala (2000), report a contrasting finding where more knowledgeable searchers mark fewer items as relevant. Differences in the methodologies used and research questions asked can give rise to different findings, as can differences in the variables being measured (the difference between knowledge of the information problem and information topic involved in a search for example). However, what these investigations and others have shown is that if we understand more about how people assess documents we gain a better insight into retrieval system success. We also gain a better understanding of what information it is useful to ask for from a searcher. That is, if we know what variables, such as knowledge on the search topic, or understanding of the purpose of a search, are useful to know about then we can ask the searcher for information regarding these variables and personalise retrieval towards the individual searcher. Understanding more about important search variables is the main aim of this paper and also one of the core aims of the TREC HARD track outlined in the next section.

3. TREC HARD track

TREC is an initiative to facilitate research on large test collections (Voorhees and Buckland, 2006). Investigations into specific retrieval tasks are organised into tracks, in which a number of participating research groups tackle the same research problem.

One of the newest of these tracks is the HARD (High Accuracy Retrieval from Documents) track (Allan, 2005) and the research described in this paper arises from our participation in the 2005 track. In this section we provide a brief overview of the HARD track to describe the context in which this research was undertaken, a fuller description of the track being given in (Allan, 2005).

The goal of the HARD track is to investigate strategies for utilising a searcher's contextual information to improve document retrieval. In other words, instead of assuming that there is an average result list that would suit all searchers, it is hoped that a ranked list of documents that is personalised to an individual searcher will improve retrieval performance. To enable this personalisation, the HARD track facilitates the capture of contextual information "by leveraging additional information about the searcher and/or the search context captured using very targeted interaction with the searcher" (Allan, 2005). The protocol for the HARD track is as follows:

(1) A number of test collections are selected for investigation. This year the track used the AQUAINT text corpus consisting of English language newswire articles from the Xinhua News Service (People's Republic of China, January 1996 to September 2000), the New York Times News Service and the Associated Press Worldstream News Service (June 1998 to September 2000 for both collections). This combined collection contains approximately one million articles.

(2) A total of 50 search topics are chosen. A TREC topic, for example:

- TREC topic 439.
- *Number*: 439 Title: inventions, scientific discoveries.
- *Description*: What new inventions or scientific discoveries have been made?
- *Narrative*: The word "new" in the description is defined as occurring in the 1990s. Documents that indicate a "recent" invention or scientific discovery are considered relevant. Discoveries made in astronomy or any scientific discoveries that are not patentable are not relevant.

Describes the criteria that will be used in assessing the relevance of retrieved documents. Each topic is assigned to a TREC assessor who will be responsible for assessing the relevance of documents retrieved for the topic; each assessor may assess more than one topic but each topic is only assessed by one assessor.

(3) Each participating group carries out an initial, baseline, run using the topics on the AQUAINT corpus. The choice of which IR system to use is up to the individual group. We selected Okapi BM25, and in particular we used the version implemented in Lemur, with the standard Lemur settings (Ogilvie and Callan, 2002). The short title of each TREC topic was used as the query, e.g. "inventions scientific discoveries" was used for the topic 439 described above.

(4) The top 1,000 ranked documents retrieved in response to each query is returned to TREC.

(5) After return of the baseline documents each group may ask the assessor questions about the topic. This is what is referred to in Allan (2005) as "by leveraging additional information about the searcher and/or the search context captured using very targeted interaction with the searcher[1]". The questions

are asked in a clarification form, described in section 4. Each participating group designs their own clarification form to gain specific pieces of information from the document assessor in order to learn more about factors that may be important in assessing the retrieved documents. As each assessor may be asked to fill in multiple forms for each topic the presentation of the forms are rotated between groups so to minimise any affect of answering one form upon subsequent forms.

(6) The information from the clarification form is used to perform a new retrieval run, e.g. by performing query expansion based on the data gathered from the clarification form, and the top 1,000 ranked documents from the new retrieval run are then returned to TREC. At this stage the assessors assess the top 100 documents retrieved for both the baseline and modified retrieval run. The results are evaluated by comparing the baseline run and the modified run to test whether the modified run improved performance for the assessor who answered the questions in the clarification form. The assessors can assess a document as being not relevant to the topic, marginally relevant or highly relevant to the topic[2].

In this paper we are not concerned with the query modification stage, which is explained separately in Baillie *et al.* (2006), Baillie and Ruthven (2006) but rather on how the assessors completed the clarification form we designed and what this tells us about the process of relevance assessment.

4. Clarification form presentation

As explained in the previous section, one of the novel features of the HARD track is the use of clarification forms by which we might gain insights into the assessors' context. Context in this sense can mean any information which might affect how the assessor determines the relevance of a document and could include information on the assessors themselves (e.g. familiarity with the topic or assessment experience), information on the information need behind the topic (e.g. why the information is required) or any question that might help decide which documents the assessor would assess as relevant. There are few restrictions[3] on the format of the clarification form except that the form must be capable of being completed within three minutes. This restriction is to control the effort placed on the assessors.

The design of the clarification form raises two research questions:

RQ1. What information is it useful to collect in a clarification form? The clarification form is intended to help us understand what type of documents an assessor would like retrieved so that a new search can be personalised to the individual searcher, in this case the TREC assessor. Although there are many pieces of information we could collect, what is useful to know about a searcher and their information need – what information helps us select better retrieved documents?

RQ2. How much useful information is it possible to obtain within three minutes? By asking for different types of information in the clarification form we can test whether some types of information are more useful than others in determining what type of information a searcher would like. This can help interface designers elicit information from searchers as part of the search process.

Three minutes is not a long time for assessors to spend on a form but it would be a long time if real searchers were asked to spend this time at an interface before being presented with any retrieved documents. Knowing what information we can obtain quickly can give insights into how IR interfaces could elicit useful information from a searcher.

The clarification form we designed is shown in Figure 1. The form is comprised of three sections each designed to gain different information from the assessors regarding their knowledge of the topic, their attitude to assessing the topic and their ability to predict which information might be relevant to the topic. These issues will be now discussed in the remainder of this section, as well as the responses given to the questions.

4.1 Topic familiarity

The first section on our form investigated topic familiarity. The degree to which an assessor is familiar with the topic of a search task, as opposed to the task itself, is interesting because we might assume that the more knowledge the assessor possesses then the more accurate the assessor can be about judging relevance and previous research, e.g. Hsieh-Yee (1993), Michel (1994), have shown that topic familiarity can affect a searcher's search strategy and search behaviour.

In this section of the clarification form we asked the assessor to assess their knowledge on the specific topic addressed in the TREC topic and on the general area described by the topic (questions 1 and 2). The questions had the form: "How much do you know about the specific topic X?" and "How much do you know about the general topic Y?" where X and Y were manually created descriptions of the specific topic and

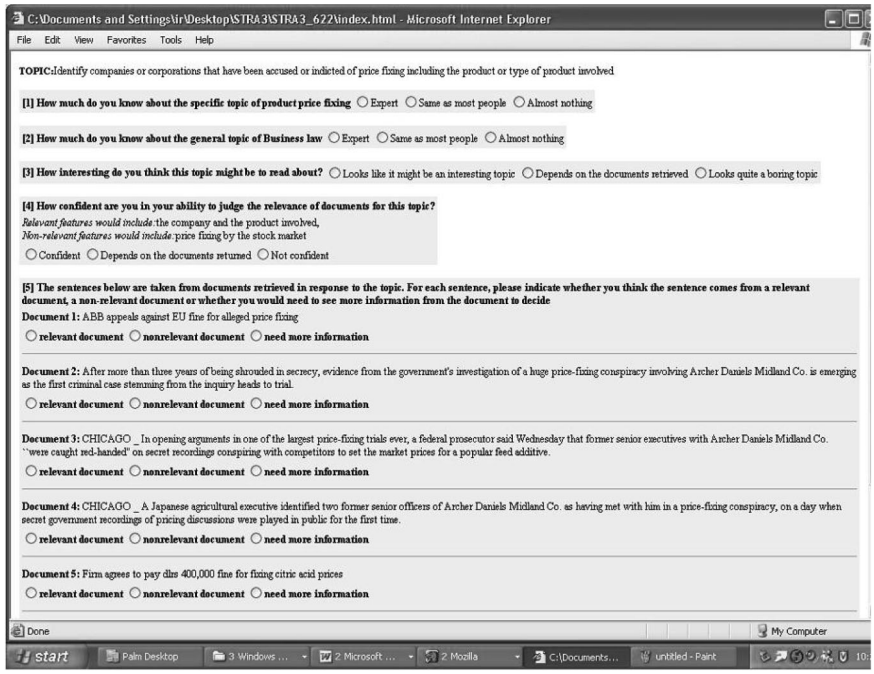

Figure 1.
The clarification form

general areas covered by the topic. In both cases the assessors were asked to rate their prior knowledge into one of three categories: "expert/same as most people/almost nothing". The rationale behind asking for both specific and general knowledge on a topic is that assessors may have good general knowledge on an area but lack knowledge on the specific sub-area that is addressed by the topic. For example, in topic 344, the assessor may know little about the specific topic of preventing the abuses of email but does know about the general topic of email, background information which can be used in searching. An assessor who has no specific or general knowledge on a topic may be seen as being less familiar than an assessor who only lacks specific topic knowledge. The combination of responses to these two questions gives nine possible combinations. In Table I we present the number of responses which fell into each combination.

The most common responses given by the assessors were to assert a medium knowledge of both the specific and general topic areas (which corresponds to the categories "[know] same as most people")[4]. For six topics the assessors claimed a higher level of specific than general knowledge but for most topics (39) the assessors gave the same rating for their knowledge of the specific and general topic. For 11 topics the subjects rated their specific topic knowledge as low ("almost nothing"), for these topics the assessors were unfamiliar with the topic before the assessment procedure began although they may have become more familiar as they read and assessed the retrieved documents.

4.2 Confidence in assessment and interest in topic

The second section on the clarification form investigated the assessor's interest and confidence in the topic. Questions 3 and 4 asked about the assessors' interest in the topic "How interesting do you think this topic might be to read about?" which was to be answered using the scale "looks like quite an interesting topic/depends on the documents retrieved/looks quite a boring topic". The second question asked the assessors to rate their existing confidence in assessing documents for the topic, "How confident are you in your ability to judge the relevance of documents on this topic?" and they were asked to rate this confidence using the scale confident/depends on the documents returned/not confident. The category "depends on the documents returned" was intended to act as a category that reflected an uncertain confidence in the assessor's ability to judge relevance or interest in the topic, the level of uncertainty being dependent on the exact nature of the documents they were asked to assess.

The reason for asking about confidence was to investigate whether an assessor's stated confidence in assessing relevance for a topic was justified: does self-declared

Specific knowledge/general knowledge	No. of responses	Specific knowledge/general knowledge	No. of responses	Specific knowledge/general knowledge	No. of responses
Low/low	7	Medium/low	6	High/low	0
Low/medium	4	Medium/medium	28	High/medium	0
Low/high	0	Medium/high	1	High/high	4

Notes: Low corresponds to the category "[know] almost nothing", medium corresponds to the category "[know] same as most people" and high corresponds to the category "expert"

Table I.
Distribution of responses to level of specific and general topic knowledge

confidence lead to more accurate assessments? The reason for asking about the assessor's interest in the topic was to uncover more information about the factors that lead assessors to make assessments of relevance: does an interest in a topic lead to a different pattern of assessments than a lack of interest?

In Table II we present the results for these questions. The assessors felt confident in their ability to judge relevance before seeing any documents retrieved for only 22 topics and unconfident for only three topics. In most cases they felt that the process of being able to judge relevance was going to be dependent on the documents actually retrieved. Although TREC topics are designed to be clear and unambiguous the assessors apparently still felt a degree of uncertainty regarding their confidence in assessment. The HARD track protocol does not allow for questioning the assessors after they have assessed the documents so we have no data on the ease of the assessment process itself, only on the assessors' predictions of how easy they felt the process would be. For most topics the assessors were either interested in the topic (30 topics) or at least could be interested depending on the documents returned. For a number of topics (seven topics) the assessors declared a lack of interest in the topic based on the topic description alone. As will be shown later there was, however, no evidence of a correlation between the assessors' interest in a topic and their confidence in assessing the relevance of documents for a topic.

4.3 Document surrogates

The final section on the clarification form was a section which presented a series of document surrogates. Specifically, question 5 presented the assessor with six fragments of text taken from the content of articles retrieved in response to the baseline query. These sentences were either headlines (Associated Press) or first paragraphs[5] (*New York Times* and *Xinhua*) and were chosen as these sentences often act as summaries of the full text of the entire newspaper article (Brandow *et al.*, 1995). In addition, the short length of these sentences meant that assessors could read and assess the sentences within the time limit of the form and sentences have been successfully used before as document surrogates, e.g. White *et al.* (2002). In this work, we examine what assessor characteristics make sentences useful as surrogates and how effective such surrogates might be to gain an insight into what a searcher will judge as relevant.

We selected the sentences to be presented to the assessor by creating two queries: a discriminatory query and a representative query. Each query was formed from a combination of the original query and either the top 12 terms that were representative of the vocabulary of the documents in the baseline run or the top 12 terms that were

Confidence in assessment	No. of topics
Confident	22
Depends on the documents retrieved	24
Not confident	3
Interest in topic	
Looks like it might be an interesting topic	30
Depends on the documents retrieved	12
Looks quite a boring topic	7

Table II.
Distribution of responses to interest in topic and confidence in assessment

discriminatory to the documents from the baseline run. Terms that are representative of the documents are those that are general terms to the topic, terms that are discriminatory are those that are very specific to the topic of the documents. This approach was inspired by the work of Harper *et al.* (2005), in the previous year's HARD track, which used the representative and discriminatory terms for retrieving documents. Here we use the expansion technique for creating a new query to retrieving sentences only. The method for choosing these terms is based on a language modelling approach explained in detail in Baillie *et al.* (2006). Table III shows the top ten representative and discriminatory terms for topic 439 (shown as an example, above).

From Table III we can see that the representative terms are ones that express more general concepts ("patent, science, human") whereas the discriminatory terms are ones that express more specialised concepts ("bioenergetics, monoclonal, thermodynamic"). In Table III we also show the comparative collection term frequency, the raw count of term occurrences for each term in the lead paragraph or first sentence, and the sentence frequency, the number of lead paragraphs or first sentences that contain each term. The representative terms generally appear in far more sentences and appear more often within each sentence. On average, over the terms chosen to expand the queries used in this study, the representative terms appeared in 90,000 sentences with an average of two occurrences within each sentence. The discriminatory terms appeared in fewer sentences (230 on average) and fewer times within each (1.5 occurrences on average). The query terms, for reference, appeared on average in nearly 28,000 sentences with an average occurrence of 1.89 times within each sentence.

The sentences were ranked according to their similarity to each of the two queries with the top three sentences taken from each ranking and presented to the assessor. The sentences chosen by the query expanded by the representative terms were intended to show the assessors sentences that contain general information on a topic. The sentences chosen by the query expanded by the discriminatory terms were intended to show the assessors sentences that contain specific information on a topic. The aim of showing these sentences to the assessors is to investigate whether the

Representative terms	Collection term frequency	Sentence frequency	Discriminatory terms	Collection term frequency	Sentence frequency
Americans	80,888	50,525	Benzinger	67	18
Cuba	35,340	12,923	Bhojwani	6	5
Discoveries	3,130	2,483	Bioenergetics	5	5
Dna	10,263	4,144	Biotechnological	37	34
Human	98,208	57,932	Biothermodynamics	4	4
Made	349,580	245,668	Csir	20	8
National	357,536	228,201	Extrasensory	35	34
Patent	6,884	2,651	Maplewood	113	86
Percent	669,084	218,234	Monoclonal	79	50
Science	51,478	32,511	Nsf	166	106
Scientific	22,637	16,052	Pares	59	51
Scientists	35,872	19,461	Preventions	17	16
Technology	112,854	63,105	Recombinant	100	73
Will	1,434,697	537,769	Thermodynamic	20	19
World	506,709	263,578	Tumbler	44	41

Table III. Representative and discriminatory terms for TREC topic 439

assessors could make predictions on what documents they would assess as being relevant before seeing the full-text of the documents.

For each sentence, the assessors were asked to judge whether the sentence came from a relevant document, non-relevant document or whether they would need to see the whole document to decide on relevance. Results from this section are presented in context in sections 5.2. and 5.3.

5. Findings

In this section we present our main findings from this study. We group the findings into three subsections; section 5.1 reports on the relationships between an assessor's knowledge, confidence and interest, sections 5.2 and 5.3. reports on the use of the first sentences as a means for predicting relevance, and section 5.4 combines these two areas of investigation. To investigate these issues we use the outputs from the TREC assessments, the final relevance assessments given to the assessed documents by the TREC assessors.

5.1 Knowledge, confidence, and interest

First we examined whether there were any relationships between the assessors' knowledge on the topic, their confidence in assessing documents on the topic and their interest in the topic. The rationale behind this is to see whether answers to one aspect of the assessors' context can substitute for another, e.g. does information about an assessors' confidence tell us something about their level of specific knowledge or are these items of information we should gather separately? In all cases we measured correlations using the non-parametric Spearman Rank test. The correlation results are presented in Table IV with statistically significant responses shown in italics[6].

For this group of assessors there was a correlation between the assessors' general and specific knowledge on the topic and a weaker correlation between their specific knowledge and interest, and their general knowledge and interest. Scores given to confidence, however, did not correlate with either knowledge or interest. So, although interest may be related to the assessor's prior knowledge of the topic, i.e. assessors know more about topics in which they are interested, we cannot assume topical knowledge or interest in a topic can be treated as substitutes for measuring confidence in assessing documents for that topic.

What we can do, however, is examine how these measures relate with the number and type of relevant documents found after the assessors have rated the documents. As mentioned before the assessors assess returned documents according to a three category scale: non-relevant, marginally relevant and highly relevant. The assessors

Table IV.
Correlations between responses to knowledge and attitude questions

Variables	Correlation values	Confidence level
Specific vs general knowledge	*0.612,* $p = 0.000$	1%
Specific knowledge vs confidence	0.141, $p = 0.339$	Not significant
Specific knowledge vs interest	*0.467,* $p = 0.001$	1%
General knowledge vs confidence	0.129, $p = 0.386$	Not significant
General knowledge vs interest	*0.331,* $p = 0.021$	5%
Confidence vs interest	-0.086, $p = 0.563$	Not significant

have no control over how many documents they assess, this number being governed by how many different documents are returned by the participating groups, but the assessors do control how they assign relevance scores to the documents and to which documents they assign a decision of marginal or high relevance. For the purposes of this paper we use four measures to analyse the pattern of relevance assessments made by the assessors:

(1) The *overall precision* which is calculated as the percentage of assessed documents that are judged to be either highly or marginally relevant. Precision is normally an absolute value associated to a search engine's performance. Here precision is calculated in the same way as standard, the proportion of retrieved (assessed) documents that are relevant. However as we demonstrate, because the assessor's context can affect judgments of relevance precision is really perceived precision: the precision of the search as perceived by an assessor with an individual personal context.

(2) The *marginal precision* which is calculated as the percentage of assessed documents that are judged to be marginally relevant

(3) The *high precision* which is calculated as the percentage of assessed documents that are judged to be highly relevant. High precision and marginal precision both contribute to overall precision; however, separating out these different types of documents allows us to see if there are any differences in how these relevance categories are used.

(4) The percentage of relevant documents assessed as being only marginally relevant (*%marginal*). This measure can be interpreted in two ways either, as in Sormunen (2002) meaning documents that are less relevant or as in Spink *et al.* (1998) meaning documents that somehow challenge the assessor's understanding of the information problem. In the following analysis we argue that the latter interpretation is probably the most appropriate one for this study.

In Table V we present the mean of these values for all topics (row average), and the mean for topics across the three responses on the questionnaires for specific knowledge, interest and confidence. This allows us to examine whether there are any affects of varying knowledge, confidence or interest on the assessors' assessments of relevance.

From Table V we can see that specific topic knowledge, confidence and interest interact differently with the assessments of relevance. More precisely:

- *Specific knowledge*. The level of specific topical knowledge[7] correlates significantly and inversely with the percentage of marginal relevant documents, %marginal, (-0.389, $p = 0.006$, 5 per cent confidence level). That is, as specific knowledge on the search topic increases a smaller proportion of the relevant documents are marked as being marginally relevant rather than highly relevant. This reinforces the findings noted by Spink *et al.* (1998) that high knowledge of a topic leads to more assessments of high relevance, while lower knowledge of a topic leads to more assessments of marginal relevance. The level of specific knowledge also correlates significantly with overall precision (0.346, $p = 0.004$, 5 per cent confidence level) and more strongly with high precision

	% marginal	Marginal precision	High precision	Overall precision
Average	61.41	10.35	7.91	18.26
Specific knowledge				
Know almost nothing (11)	83.45	8.01	1.91	9.92
Same as most people (35)	57.27	11.45	9.06	20.51
Expert (4)	36.98	7.21	14.31	21.52
Confidence				
Not confident (3)	78.20	8.06	2.22	10.27
Depends on the documents returned (24)	44.28	7.52	9.46	16.98
Confident (22)	71.85	13.88	6.92	20.80
Interest				
Looks quite a boring topic (7)	92.23	8.05	0.77	8.82
Depends on the documents retrieved (30)	49.43	6.88	8.67	15.55
Looks like it might be an interesting topic (12)	59.41	12.39	9.24	21.64

Notes: Figures in parentheses are the number of assessors in each category

Table V.
Relevance assessments for high/medium/low levels of knowledge, confidence and interest

(0.402, $p = 0.004$). Higher precision is unlikely to arise because there are more relevant documents for the topics with which the assessors are familiar. If this were the case we would expect the number of marginally relevant documents to increase at the same time. Rather, it appears that when assessors know more about the topic which they are assessing, they mark more of the retrieved documents as relevant and more of them as highly relevant rather than marginally relevant.

- As *confidence* in assessing relevance increases[8] the overall precision increases although there is no significant correlation between overall precision and confidence score. The reason for the increase in precision at high confidence levels is the high value attached to marginal precision, the percentage of assessed documents marked as marginally relevant. For low levels of confidence the percentage of retrieved documents marked as highly relevant is very small (2.2 documents per topic) suggesting a reticence to commit to high relevance decisions.

  The %marginal score is almost identical for the situations where the assessors are confident or not confident, i.e. where the assessors give definite scores for their confidence before seeing any documents. For these cases there are more documents being marked as marginally relevant and fewer as highly relevant. What is interesting about these results is that the middle option to the question ("depends on documents retrieved") is not a middle range. That is, for most of the evaluation measures the middle option to the confidence question does not behave as an option reflecting a middle way between a definite sense of confidence or a lack of confidence in ability to judge relevance. For example, the values for %marginal and marginal precision are lower than for the other two options and the value for high precision is higher. In this case the middle option appears to be employed by conservative assessors: ones who are less willing to predict their ability to judge relevance, and less willing to use marginal

relevance. Why this occurs, what decisions are being made here, we cannot directly answer as we do not have access to the TREC assessors, but we point to it as an area for further investigation with end users.

- As *interest* in the topic of the search task increases[9] both high precision and overall precision increase. This relationship was found to be significant for both high and overall precision measures (0.346, $p = 0.015$ for overall precision, 0.304, $p = 0.034$ high precision, both at 5 per cent confidence level only). As with confidence the role of the middle option is not straightforward and, like the responses to confidence, the middle option has low %marginal and low marginal precision suggesting a low use of the marginal relevance category and conservative assessment behaviour.

  If we exclude responses given to the middle option (the responses in the category "depends on the documents retrieved") and concentrate only on the topics for which the assessors expressed a positive interest or lack of interest then interest still correlates with overall precision (0.407, $p = 0.012$, 5 per cent confidence level), more strongly than previously with high precision (0.576, $p = 0.000$) and also with %marginal (-0.534, $p = 0.001$). The correlation with %marginal is negative meaning that increased interest in a search topic correlates with a smaller percentage of relevant documents being marked only marginally relevant and a higher percentage of relevant documents being marked highly relevant. Together these correlations mean that there is strong evidence of a positive relationship between interest and the number of relevant documents highly relevant. As interest increases the preference is to mark a relevant document highly relevant rather than marginally relevant.

  Interest itself may affect the assessors judgments either by the assessors being more willing to consider a document relevant if they are interested in the topic, by the assessors taking more time to read documents carefully if they are interested in the topic or by the assessors seeing more connections between documents and the topic if they are interested in the topic. We are currently investigating these questions in new user tests.

The different searcher characteristics, then, all have an effect on the final precision of the search even though this relationship is not straightforward when we examine how they use the marginal relevance category. These are the decisions made on the documents after assessment. In the following sections we look at whether these factors also affect decisions made before seeing the documents.

5.2 Quality of sentences as predictors of relevance

The final section of the clarification form asked assessors to predict the relevance of documents based on short headline sentences or first paragraphs. In an interface this technique might be useful to help searchers decide which documents to read in more detail or to gain more knowledge on what information a searcher wants to read (White *et al.*, 2002). The sentences may also be seen as another way of estimating searcher attributes. In the form we designed, for example, the number of sentences for which an assessor was willing to make a relevance prediction (relevant/not relevant) could be considered as an indirect estimate of their confidence in assessing relevance. For the sentences to be useful, however, we first must assess how useful they are at helping

assessors predict relevance and how stable these assessments are with respect to varying interest, confidence and knowledge.

In Table VI we present (column 2) the percentage of sentences that were predicted as coming from a relevant document, coming from a non-relevant document or where the assessor felt they needed more information to judge relevance. In column 4 ("column % of documents assessed in each category") we present the percentage of sentences that came from a document that was finally assessed as being relevant or non-relevant.

Firstly, in examining the assessors' confidence in judging the sentences we find that the assessors felt that they could make a prediction on the relevance or non-relevance of the documents based only the headline or first sentence in the majority of cases (71 per cent of sentences chosen using representative terms and 62 per cent of sentences chosen using discriminatory terms[10]), Table VI. For both sets of sentences the assessors predicted non relevance more often than they predicted relevance. However, as approximately 35 per cent of sentences belong to documents that were finally assessed as being relevant, predicting fewer sentences as coming from relevant documents is not necessarily a reflection on the quality of the sentences.

As noted in the previous paragraph, the assessors made more predictions of relevance for the sentences chosen by the representative terms, making a prediction of (non)relevance for 71 per cent of sentences presented against only 62 per cent of sentences chosen by the discriminatory terms. For the sentences chosen using discriminatory terms the assessors were also far less likely to predict relevance than for the sentences chosen by the representative terms. Harper *et al.*'s results showed expansion by discriminatory terms benefited document retrieval (Harper *et al.*, 2005). Our results indicate that representative term expansion may benefit interactive retrieval by selecting good sentences, i.e. those for which assessors see as potentially relevant to their search.

There was no evidence of any correlation between the assessors' responses to the topic/confidence/interest questions and their responses in this part of the form. This lack of correlation with confidence in particular is interesting because the assessors generally make confident predictions on relevance, feeling confident enough to predict relevance on up to 70 per cent of sentences presented.

Table VI.
Confidence in predicting relevance based on headline or first paragraphs

	% of sentences assigned to each category	Final assessments on documents	% of documents assessed in each category
Predictions on sentences chosen by discriminatory terms			
Came from relevant document	17	Relevant	35
Need more information	38	–	–
Came from non-relevant document	45	Not relevant	65
Predictions on sentences chosen by representative terms			
Came from relevant document	31	Relevant	37
Need more information	29	–	–
Came from non-relevant document	40	Not relevant	63

However, if we compare the number of sentences for which the assessor was willing to provide a definite prediction (relevant or not-relevant) with the end results of the assessment process, then we do find relationships. In particular we find that the number of sentences for which the assessor makes a confident prediction correlates negatively with high precision (the number of sentences assessed relevant/not relevant against high precision, -0.310, $p = 0.028$, 5 per cent confidence level). This means that if an assessor marks many sentences as relevant or not relevant then they will end up marking fewer documents as highly relevant. The total number of sentences predicted as belonging to a relevant/not-relevant documents also correlates with the %marginal (0.320, $p = 0.023$, 5 per cent confidence level). This implies that an assessor's willingness to predict relevance could be related to how conservative the assessor will be in assessing relevance: if an assessor is unwilling to predict relevance before seeing the full-text of a document they are less likely to use the marginal relevance category and more likely to mark a document highly relevant: a pattern that fits the conservative assessor described in section 5.1.

5.3 Predictions vs outcomes

The predictions on sentences are predictions made before the assessors see any documents. In Table VII we compare the accuracy of these predictions by examining the assessments made after reading the documents (not relevant, marginally relevant, highly relevant) and the assessors' predictions on the clarification forms (not relevant, cannot tell, relevant). What we are comparing here is, of course, the assessors' predictions on what relevance decisions they will make when they see the full-text of the documents. The combination of these predictions and final assessments gives nine categories and, in Table VII, we present the percentage of sentences that fall into each category for the assessors.

Numerically, the most common pattern was the correct identification of non-relevance (non-relevant/non-relevant) into which fell approximately 37 per cent of sentences, although as most documents were non-relevant we might expect this pattern to be common. The next trend was an assessor's lack of certainty on predicting the relevance of a document which was finally assessed as being non-relevant (category non-relevant/need more), into which fell 19 per cent of sentences.

In the cases where the assessors made a confident (relevant/non-relevant) prediction we can define two types of accurate predictions[11]:

Prediction/assessment	%	Prediction/assessment	%	Prediction/assessment	%
Non-relevant/non-relevant	36.63	Need more information/non-relevant	19.31	Relevant/non-relevant	9.41
Non-relevant/marginal	4.95	Need more information/marginal	8.42	Relevant/marginal	9.9
Non-relevant/high	0.99	Need more information/high	5.94	Relevant/high	4.5

Notes: Assessment is the assessors' relevance assessment after reading the document and prediction is their prediction on the document based on the headline or first paragraph

Table VII.
Accuracy of sentences in predicting relevance

(1) Accurate predictions on relevance, i.e. predictions that the sentence was taken from a relevant document when the document was finally assessed as being either highly or marginally relevant (relevant/marginal and relevant/high categories). Of the total sentences predicted to be relevant the assessors were correct in 61 per cent of cases, with most of the inaccuracy coming from the assessors over-estimating the presence of relevant information (category relevant/non-relevant).

(2) Accurate predictions on non-relevance, i.e. that a sentence was taken from a non-relevant document when the documents was finally assessed as being non-relevant (non-relevant/non-relevant). Of the total sentences predicted to be non-relevant the assessors were correct in 86 per cent of cases.

If we examine only the cases where the assessors could make a prediction then we find that for about 51 per cent of sentences presented the assessors could make a prediction, and made the correct prediction. On the other hand, for approximately 15 per cent of sentences the assessors could make a prediction but made the incorrect prediction. So the assessors are more likely to make a correct prediction than an incorrect prediction, if they make a prediction at all, but there is room for improvement in presenting better surrogates: ones that allow assessors to make more relevance predictions and make more correct predictions.

As we show in the next section, the assessors' ability to predict relevance accurately is not solely based on the quality of the sentences but also on the assessors' relationship with the search topic and process of assessment.

5.4 Effects of knowledge, confidence and interest on quality of sentences as predictors of relevance

In the previous section we showed that assessors could only make accurate predictions on relevance, based on sentence evidence, for about half the sentences presented. In Table VIII, we break down this prediction accuracy figures to show how confidence, interest and knowledge affect these predictions. Table VIII shows:

- The percentage of sentences displayed on the form for which the assessor made a prediction (column 2). This column indicates the relationship between the assessor attributes and predictions.
- The percentage of predictions that were relevance predictions, i.e. assessor made a prediction that a sentence was relevant (column 3).
- The percentage of sentences that came from documents finally assessed as being relevant (column 4). This is the percentage of sentences upon which the assessor made a prediction of relevance. This column therefore shows how many sentences, on average, should be predicted as relevant if the assessment process if not affected by the assessor context.
- The accuracy of the predictions against the assessors' final assessment of the document containing the sentence (column 5 for accuracy of relevance predictions, column 6 for accuracy of non-relevance predictions).

For *specific topical knowledge*, as the assessors' topical knowledge increases they make predictions on more sentences, more predictions of relevance and fewer predictions of non-relevance. Only assessors with a high declared level of topical knowledge in this study predicted relevance more often than they predicted non-relevance, although most

	Percentage of sentences with prediction	Percentage of sentences predicted as relevant	Percentage of sentences from relevant documents	Accuracy of relevance predictions (%)	Accuracy of non-relevance predictions (%)
Knowledge					
Know almost nothing (11)	64	31	28	70	91
Same as most people (35)	66	35	31	63	85
Expert (4)	77	60	30	33	75
Confidence					
Not confident (3)	50	66	50	75	100
Depends on the documents returned (24)	68	26	15	32	91
Confident (22)	67	45	48	80	77
Interest					
Looks quite a boring topic (7)	51	66	60	80	80
Depends on the documents retrieved (30)	60	52	38	56	80
Looks like it might be an interesting topic (12)	73	25	22	55	89

Table VIII. Accuracy of sentences in predicting relevance for differing levels of confidence, knowledge and interest

documents were finally assessed as being non-relevant. However, at the same time, increased topical knowledge led to a fall in the assessors' ability to predict their final decision, both for relevance and non-relevance. That is, with high topical knowledge, the assessors were less accurate in predicting their final relevance decision. What is happening here is that the assessors are over-estimating the presence of relevance in the case where they report high topical knowledge: predicting too many documents will be relevant and too few documents will be non-relevant. As with the final relevance assessments, section 5.1, in the situation where the assessor has high topical knowledge it is not the case that there are more relevant documents rather that the assessor is more willing to assign a decision of relevance.

For *confidence* we see a different pattern. If we compare the cases where the assessors make a definite assessment of their confidence (not confident/confident) then, as assessors' confidence increases, they become more willing to predict non-relevance for a sentence and less willing to predict relevance. As before, the responses to the middle option to the question ("depends on the documents retrieved") fit the pattern of assessors who are behaving conservatively towards relevance: predicting most sentences come from non-relevant documents and few come from relevant documents. For this case the percentage of sentences that came from relevant documents was very low (15 per cent compared to 48 per cent/50 per cent for the other categories). Even though the assessors in this category made few predictions on relevance, the low percentage of sentences from relevant documents means that the assessors are still over-estimating relevance and have a low value for the accuracy of their relevance predictions. The low value for relevant sentences in this category does not mean that there were few relevant documents. As

shown in Table V, this is not the case. What it does mean is that either the documents that the assessors would finally mark as relevant were not typical of the ones presented on the clarification form, or the sentences were not ones that could allow the assessors to make good predictions.

For *interest*, we see an opposite pattern to that for specific knowledge with increased interest being associated with fewer predictions of relevance and more predictions of non-relevance. However, although the assessors are good at predicting non-relevance, they become less able to predict what they will finally assess as being relevant. As in the previous cases, the inaccuracy in relevance predictions comes from over-estimating relevance information.

This section examined whether sentence level information could be useful in detecting relevance information from a searcher. In general the method we used, selecting first paragraphs or first sentences, worked relatively well with the assessors being able to make predictions for the majority of sentences and making correct predictions more often than incorrect predictions. However there is still room for improvement to better these results and we showed that these figures are affected by the searchers' contextual factors. In particular more knowledge and interest in a topic appeared to lead the assessor to make more predictions and over-estimate the presence of relevance information. Over-estimating the presence of relevant information may not be a poor feature if it leads a searcher to interact more with search results, and is preferable to over-estimating non-relevant information which may lead the searcher to miss important information. We still need a good explanation for the role of declared confidence especially what interactive support is required for searchers in the middle category (who answered "depends on the documents retrieved" to the confidence question) as these searchers are making different decisions than those who could give a definite decision on their confidence.

6. Discussion

What our study shows is that a searcher's personal context, their knowledge in and attitudes to a search task, affect how they assess relevance. Our results indicate that assessors with high self-declared knowledge regarding a search topic, high interest in the search topic or high confidence in assessing documents will assess more documents as relevant to a search than searchers with lower topical knowledge, interest and confidence. The implication of these results is that we need to look more closely at how assessors' personal characteristics interact with relevance decisions, especially for the evaluation of IR systems. The difference, for example, between the average precision for low interest/confidence/knowledge topics is half that for searches where the assessor has high interest, confidence or knowledge on the topic. Such differences in assessment behaviour must make a difference to how well we understand the performance of IR systems as well as creating search tasks. For creating search tasks we should consider not only the searchers' topical knowledge, as is commonly measured, but also the searchers' interest in and confidence in the search topic.

We also asked what information it is useful to gather in a search interface. We presented two methods of gaining information: asking direct questions and presenting sentences as information surrogates. Presenting sentences could act as an implicit method of gaining information on a searchers' context rather than asking direct questions. On average, the sentences were useful but we did show that the assessors' characteristics could make the sentences more or less useful depending on the assessors' ability to predict what they would find relevant and their willingness to

predict relevance based on sentence evidence. In different search scenarios we should consider what interactive support to provide for individual searcher as one method of creating surrogates may not be appropriate for all situations.

One specific question we asked was whether self-declared confidence, i.e. responses to the question "How confident are you in your ability to judge relevance", leads to more accurate assessments? The answer from this study is no: assessors who were willing to make a declaration on their confidence level behaved similarly in some respects (higher use of marginal precision, similar percent of documents marked as marginally relevant, good ability to predict their final relevance assessments) but very differently from those assessors who could not declare their confidence or, at least, felt their confidence would be dependent on the documents retrieved. These assessors, who we classified as conservative, are interesting in that their response to the confidence question highlights a different type of behaviour rather than a different level of behaviour, i.e. except for the overall precision scores their relevance assessment behaviour was very different from the rest of the assessors' behaviour. Why this is the case is something we are investigating currently as we believe it may help us understand better what questions to ask searchers to understand what information they require. However, this finding suggests that we need to look closely at what information we gather from searchers and how we use this information.

7. Limitations of study

Our study is limited in several respects, mostly arising from the construction of the design of the study. One major limitation of our study, for example, is that it was performed within the context of the HARD track and therefore we are limited in only being able to deal with the outputs of the relevance assessment process and cannot ask for elucidation from the assessors on their assessment decisions. Neither did we ask for more information on, for example why an assessor might have felt confidence or lack of confidence in assessing relevance or why they might have been interested in reading about a topic. The questions asked are also themselves contextually dependent; if we had asked the questions at different points or for different task constructions we might have obtained different responses.

As the variables of concern (topical knowledge, interest and confidence) were not part of the official TREC protocol we were not able to select assessors or select topics for individual assessors. Therefore, the numbers of assessors in each category were very different and, for some groups, were very small.

A final limitation, and one noted earlier, was that we are dealing with assessments rather than interaction and as a result our analysis should not be confused with the results of interactive searching.

8. Conclusion

This is one study and the results will need inspection against further studies. Nevertheless the results we obtained indicate that the three variables we investigated share some common properties. For example, they all affect the precision of a search. That is, the values of these variables affect how people, in this case TREC assessors, judge relevance for a search. As the TREC HARD track allowed assessors to mark documents as marginally relevant as well as highly relevant we were able to distinguish

how these variables affected the balance between marginal and high relevance decisions that would not have been obvious with only binary relevance decisions.

Our results particularly point to the fact that we need to consider carefully how we create search tasks for IR evaluations. As we have shown in this paper, a searcher's personal knowledge of a search topic and their attitude to the search topic relates to the assessments they make on documents retrieved. By comparing knowledge and attitude variables within the same study we have demonstrated that these variables are linked but are not directly substitutable; high levels of interest, for example, were linked with fewer predictions of relevance but high levels of knowledge were associated with more predictions of relevance. Gathering more information on different aspects of the searchers' context would appear to be beneficial to allow interactive IR systems to gain a better understanding of what information a searcher might require and what decisions a searcher might require support.

Notes

1. In the context of the Hard Track the searchers are TREC assessors, non-IR specialists who are responsible for assessing documents returned. We will continue to use the term assessors in the remainder of this paper to avoid the confusion that we are interacting with end users of a system: the TREC assessors in the HARD track see a number of clarification forms and documents retrieved by different IR systems but are not interacting with any IR system.
2. The HARD track refers to these categories as non-relevant, relevant and highly relevant. We use marginally relevant in place of relevant for clarity.
3. There are some technical restrictions that are not relevant to this paper, e.g. the forms cannot invoke cgi scripts or write to local disks.
4. For a small number of questions (eight out of the 500 questions in total) there were no responses made on the clarification form, i.e. no radio button clicked. We present the results only for questions answered.
5. First paragraphs were chosen as these collections did not have a headline field.
6. Unless stated, all significant correlations are measured at the 0.001 (1 per cent) confidence level (two-tailed).
7. Know almost nothing/same as most people/expert.
8. Not confident/depends on the documents returned/confident.
9. Looks quite a boring topic/depends on the documents retrieved/looks like it might be an interesting topic.
10. Summing the values in categories came from relevant document and came from non-relevant document.
11. False predictions on relevance (predicting relevance on a non-relevant document) and false predictions on non-relevance (predicting non-relevance on a marginally or highly relevant document) are the complement of these two cases.

References

Allan, J. (2005), "Hard track overview in TREC 2004: high accuracy retrieval from documents", *Proceedings of 13th Text Retrieval Conference (TREC-13), Gaithersburg, 16-19 November*, NIST Special Publication: SP 500-261, NIST, Gaithersburg, MD.

Baillie, M. and Ruthven, I. (2006), "Examining assessor attributes at Hard 2005", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in IR, Seattle. 6-11 August*, ACM Press, New York, NY.

Baillie, M., Elsweiler, D., Nicol, E., Ruthven, I., Sweeney, S., Yakici, M., Crestani, F. and Landoni, M. (2006), "University of Strathclyde: the i-lab's first big day out at TREC HARD", in Voorhees, E.M. and Buckland, L.P. (Eds), *Proceedings of the 14th Text Retrieval Conference (TREC-14), Gaithersburg. 15-18 November*, NIST Special Publication no. 500-266, NIST, Gaithersburg, MD.

Barry, C.L. and Schamber, L. (1998), "Users' criteria for relevance evaluation: a cross-situational comparison", *Information Processing and Management*, Vol. 34 Nos 2/3, pp. 219-36.

Borlund, P. (2003), "The concept of relevance in IR", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 10, pp. 913-25.

Brandow, R., Mitze, K. and Rau, L.F. (1995), "Automatic condensation of electronic publications by sentence selection", *Information Processing and Management*, Vol. 31 No. 5, pp. 675-85.

Buckley, C. and Voorhees, E.M. (2005), "Retrieval systems evaluation", in Voorhees, E.M. and Harman, D.K. (Eds), *TREC Experiment and Evaluation in Information Retrieval*, MIT Press, Cambridge, MA, pp. 53-78).

Choi, Y. and Rasumussen, E.M. (2002), "Users' relevance criteria in image retrieval in American history", *Information Processing & Management*, Vol. 38 No. 5, pp. 695-726.

Eisenberg, M. and Barry, C. (1998), "Order effects: a study of the possible influence of presentation order on user judgements of document relevance", *Journal of the American Society of Information Science.*, Vol. 39 No. 5, pp. 293-300.

Eisenberg, M. and Hu, X. (1987), "Dichotomous relevance judgments and the evaluation of information systems", *Proceedings of the American Society for Information Science 50th Annual Meeting, Boston, 4-8 October*, Learned Information, Medford, NJ, pp. 66-9.

Florance, V. and Marchionini, G. (1995), "Information processing in the context of medical care", in Fox, E.A., Ingwerson, P. and Fide, R. (Eds), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in IR. Seattle. 9-13 July*, ACM Press, New York, NY, pp. 158-63.

Harper, D.J., Muresan, G., Liu, B., Koychev, D., Wettschereck, D. and Wiratunga, N. (2005), "The Robert Gordon University's Hard Track experiments at TREC 2004", *Proceedings of 13th Text Retrieval Conference (TREC-13), Gaithersburg, 16-19 November*, NIST Special Publication: SP 500-261, NIST, Gaithersburg, MD.

Hsieh-Yee, I. (1993), "Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers", *Journal of the American Society for Information Science*, Vol. 44 No. 3, pp. 161-74.

Huang, M.-H. and Wang, H.-Y. (2004), "The influence of document presentation order and number of documents judged on users' judgements of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 11, pp. 970-9.

Järvelin, K. and Kekäläinen, J. (2000), "IR evaluation methods for retrieving highly relevant documents", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in IR, Athens, 24-28 July*, ACM Press, New York, NY, pp. 41-8.

Katter, R.V. (1968), "The influence of scale form on relevance judgments", *Information Storage and Retrieval.*, Vol. 4 No. 1, pp. 1-11.

Lesk, M.E. and Salton, G. (1969), "Relevance assessments and retrieval system evaluation", *Information Storage and Retrieval*, Vol. 4, pp. 343-59.

Michel, D. (1994), "What is used during cognitive processing in information retrieval and library searching? Eleven sources of search information", *Journal of the American Society for Information Science and Technology*, Vol. 45 No. 7, pp. 498-514.

Mizzaro, S. (1997), "Relevance: the whole history", *Journal of the American Society for Information Science and Technology*, Vol. 48 No. 9, pp. 810-32.

Mizzaro, S. (1999), "Measuring the agreement among relevance judges", in Draper, S.W., Dunlop, M.D., Ruthven, I. and Rijsbergen, C.J. (Eds), *Proceedings of the Final Mira Workshop (Mira '99), Glasgow, 14-16 April*, published in *Electronic Workshops in Computing*, April, pp. 1-13.

Ogilvie, P. and Callan, J.P. (2002), "Experiments using the lemur toolkit", *Proceedings of 10th Text Retrieval Conference (TREC-10), Gaithersburg, 13-16 November*, NIST Special Publication: SP 500-250, NIST, Gaithersburg, MD, pp. 103-8.

Ruthven, I. (2005), "Integrating approaches to relevance", in Spink, A. and Cole, C. (Eds), *New Directions in Cognitive Information Retrieval*, Information Retrieval Series, Vol. 19, Springer, Dordrecht, pp. 61-80.

Sormunen, E. (2002), "Liberal relevance criteria of TREC – counting on negligible documents?", *Proceedings of the 25th Annual International Conference on Research and Development in IR, Tampere. 11-15 August*, ACM Press, New York, NY, pp. 324-30.

Spink, A., Greisdorf, H. and Bateman, J. (1998), "From highly relevant to not relevant: examining different regions of relevance", *Information Processing and Management*, Vol. 34 No. 5, pp. 599-621.

Tiamiyu, M.A. and Ajiferuke, I.Y. (1988), "A total relevance and document interaction effects model for the evaluation of IR processes", *Information Processing and Management*, Vol. 24 No. 4, pp. 391-404.

Tombros, A. and Crestani, F. (2000), "Users' perception of relevance of spoken documents", *Journal of the American Society for Information Science*, Vol. 51 No. 9, pp. 929-39.

Tombros, A., Ruthven, I. and Jose, J.M. (2005), "How users assess web pages for information-seeking", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 5, pp. 327-44.

Vakkari, P. and Hakala, N. (2000), "Changes in relevance criteria and problem stages in task performance", *Journal of Documentation*, Vol. 56 No. 5, pp. 540-62.

Vakkari, P. and Sormunen, S. (2004), "The influence of relevance levels on the effectiveness of interactive information retrieval", *Journal of the American Society for Information Science*, Vol. 55 No. 11, pp. 963-9.

Voorhees, E.M. (2001), "Evaluation by highly relevant documents", *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, 10-12 September*, ACM Press, New York, NY, pp. 74-82.

Voorhees, E.M. and Buckland, L.P. (Eds) (2006), *Proceedings of the Fourteenth Text Retrieval Conference (TREC 2005)*, NIST, Gaithersburg, MD.

White, R.W., Ruthven, I. and Jose, J.M. (2002), "Finding relevant documents using top-ranking sentences: an evaluation of two alternative schemes", *Proceedings of the 25th Annual International Conference on Research and Development in IR, Tampere, 11-15 August*, ACM Press, New York, NY, pp. 57-64.

Corresponding author
Ian Ruthven can be contacted at: ir@cis.strath.ac.uk