Relevance behaviour in TREC

1. Introduction

Relevance assessment is a core area of interest for Information Science. Assessing a document as relevant to an information need can be a complex decision that may be affected by the searcher's knowledge, the task being conducted, the collection being searched, the retrieval results and properties of the document itself (Barry and Schamber, 1998; Harter, 1996; Ruthven, 2005; Saracevic, 2007a; Saracevic, 2007b; Spink et al., 1998; Vakkari and Hakala, 2000; Voorhees, 2001). Analysing relevance assessments can help us understand how people make judgments about relevance, how people solve information problems and, through collection of assessments within test collections, help us evaluate the performance of retrieval systems.

TREC assessors over a period of 20 years have engaged in a vast number of relevance judgments on different topics, languages, document types and retrieval tasks (Voorhees and Harman, 2005). Due to this intensive effort, the research community has been able to create a set of invaluable tools for the evaluation of new retrieval systems. However, this collection of assessments has told us relatively little about the factors affecting the judgments of relevance. In spite of being the single biggest collection of relevance judgments - aside from implicit assessments in the form of click data from search engines - there are few analyses of the human aspects of relevance assessment within TREC. Even the interactive tracks, such as the interactive track, ciQA and HARD tracks, have focused on using interaction to improve retrieval effectiveness rather than understanding relevance itself.

In this paper we try to examine how various types of TREC data can be used to better understand relevance and serve as test-bed for exploring relevance. We propose that there are many interesting studies that can be performed on the TREC data collections that are not directly related to evaluating systems but to learning more about human judgments of relevance and that this data can help uncover useful research questions for other types of investigation. One key advantage of the TREC data is its large size which allows for statistical generalisation. Such generalisation can help pose new research questions to be investigated within more qualitative studies. What we try to show in this paper is that TREC data can be a useful source of data to understand, as well as simply measure, relevance and through proof-of-concept case studies, that various types of investigation are possible to exploit the investment in TREC as a source of new relevance studies.

The paper is structured as follows: firstly in section 2 we outline some of the investigations into relevance assessment in TREC, then in section 3 we present three case studies on relevance behaviour in TREC, and in section 4 we discuss findings and future directions.

2. Related work

In his seminal paper, Saracevic described '*manifestations of relevance'*, a series of 5 relationships which might be used to describe relevance (Saracevic, 1996). These include relations between the goals, motivations and situations of a searcher and the texts with which they are interacting. Importantly, Saracevic claimed that these manifestations were a series of interdependent and dynamic system of relevance judgment rather than a discrete, uniform process of assessment. That is, context can change assessments and the effects of context can be observed within the types of relevance decision being made.

In information seeking, many researchers have investigated the reasons why people judge an information object as being relevant. A core, and repeated, finding is that topical relevance, the type of relevance commonly investigated in TREC, is usually necessary but not sufficient for a judgment of relevance (Voorhees and Harman, 2005). That is, a document which is not topically relevant to an information need is unlikely to be marked relevant but simply being topically relevant is not always enough to make a document relevant. Rather, people bring other factors to bear on the relevance judgment process as well as topical relevance (Barry and Schamber, 1998).

Since these influential studies many researchers, e.g. (Bateman, 1998; Choi and Rasmussen, 2002; Sormunen, 2002; Vakkari and Hakala, 2000), have investigated relevance within a variety of search situations (Saracevic, 2007a). These investigations, like many user studies, have typically involved relatively small numbers of searchers making small numbers of relevance judgments and often in realistic settings where the participants have fewer restrictions on how to assess relevance. TREC tracks, on the other hand, whilst still employing relatively small numbers of assessors on each track, have engineered thousands of relevance decisions within a much more controlled setting.

TREC is often criticised for preferring experimental control over realism when dealing with relevance. It is not fair to criticise TREC for being something that it is not: it was intended to be a mechanism for creating test collections and these tools demand repeatability which emphasises control. It is not clear, however, to what extent the experimental rigour of TREC impinges on the realism of assessment behaviour and the question does arise of whether similar relevance manifestations appear in TREC assessments as in user studies? In spite of robust descriptions of contextual factors which affect relevance decisions in other areas, e.g. in Web searching (Ford et al. 2001), we still lack similar understandings for initiatives such as TREC which might offer different findings on relevance behaviour.

Harter (1996) specifically criticised the standard test collection model of evaluation because it ignored the variation in these contextual factors as to why relevance judgments may be made. His major argument is that variations in the factors involved in assessment are smoothed out by the aggregation involved in test collection methodologies meaning that we lose important results about the success of individual queries and that we need to develop evaluation measures that are more sensitive to differences in why relevance assessments occurred. In this paper we try to show that studies that unpick these variations are possible, to an extent, with the TREC data.

There have already been several important studies on relevance behaviour using TREC data. Turpin and Hersh's research on why relative differences in system performance do not always lead to the same differences in interactive situations, for example, has led to a number of attempts to understand why differences in relevance behaviour that appear in interactive searching are different from those in laboratory scenarios (Kelly et al., 2010; Smith and Kantor, 2008; Smucker and Jethani, 2010; Turpin and Hersh, 2001).

A second body of work examined issues of reliability of assessment. Voorhees examined the consistency of relevance assessments, the degree to which different assessors would consistently rate the same documents as being relevant, showing that assessments for a topic could vary significantly between topics and positing an upper bound on system performance based on the level human agreement on relevance of 65% (Voorhees, 2000). Later she noted, when examining decision making on highly relevant documents, that '*Assessors frequently*

disagreed on which document was best, and the relative effectiveness of systems when evaluated by different assessors changed markedly', (Voorhees, 2001). Smucker and Jethani proposed that NIST assessors were more reliable in detecting relevance than non-TREC assessors and, intriguingly, that time taken to assess documents may indicate error in assessment (Smucker and Jethani, 2011). Other attempts to measure assessors' lack of consistency include Scholer et al.'s investigation which examined irregularities in assessment of duplicate documents within the TREC assessment procedures, proposing reasons for the lack of consistency including '*assessors were not always clear on the criteria used to judge a document and that such criteria were either forgotten, or alternatively that an assessor's view of what constituted relevance shifted over time as the documents were judged*', a situation that often happens in more naturalistic situations. Lin and Zhang examined intra-assessor consistency within the TREC Complex Interactive Question Answering (ciQA) track showing that, although assessors could be inconsistent when assessing the same answer multiple times, the error rate was not sufficient to change system orderings. They also made the useful point that *'inter-assessor differences might be more pertinent than that inter-run differences.'* (Lin and Zhang, 2007).

Most of the above research has been conducted outside the main TREC studies, creating new relevance assessments rather than studying existing data, or has focussed on the impact of differing relevance assessments on system performance rather than trying to understand what kinds of relevance judgments appear within TREC. There are exceptions. Sormunen [24], although using external assessors, examined whether we could detect different levels of relevance within TREC assessments, showing that many relevant documents were only marginally relevant to the topic. Ruthven et al. as part of the High Accuracy Retrieval of Documents (HARD) track, noted interesting relationships between assessor characteristics and relevance judgments (Ruthven et al., 2007). As assessor knowledge in a topic increases, for example, assessors are likely to mark more documents as relevant and more likely to mark documents as highly relevant. Chu examined relevance criteria use within the 2007 TREC legal track, demonstrating the importance of topicality in making assessments but also a range of other factors that could influence the assessment process (Chu, 2007). So, although TREC may be seen to be unrealistic in some ways, e.g. the use of fixed narrative descriptions, it is not clear that these lead into unrealistic relevance behaviour.

In the remainder of the paper we present three cases studies where we use publicly available TREC data to learn more about relevance behaviour within TREC. Our aim is to demonstrate how the large-scale data within TREC can lead to useful new findings that were not part of the original TREC tracks and motivate the use of TREC as a source of new research activity to better understand relevance.

3. Case studies

In this section we present three case studies to illustrate how various sources of TREC data may be used to understand relevance. We have chosen a case study approach to illustrate what is possible with TREC assessments rather than to answer the question '*What is relevance behaviour in TREC*'. Such a question is too large and would force a generalisation over too many variables, track designs and assessment tasks. Rather we hope to show, in three different studies, that different types of available TREC data can be used creatively to investigate useful relevance questions.

The three studies will focus on three factors that may affect assessments: the assessor, the topic being searched and the document collection which provides the material to be assessed.

The first examines whether assessors' characteristics lead to different relevance assessments section 3.1; the second examines the consistency of expansion term selection in the TREC-2005 HARD track, an example of where TREC data can be used to investigate interactive retrieval decisions, section 3.2, and the final study looks at whether the nature of the documents being assessed leads to different relevance decisions, section 3.3.

3.1. Assessor factors and assessment outcomes

It is well demonstrated in information seeking research that subjective factors, such as confidence, topical knowledge and personality, can affect search behaviour and search outcomes, e.g. Gwizdka and Lopatovska (2009). Some subjective factors will affect how a search is conducted, including strategies for completing a search; other factors may affect how relevance is assessed. Sormunen, for example demonstrated the use of different relevance levels between assessors (Sormunen 2002), several authors (e.g. Florance and Marchionini, 1995, Eisenberg and Barry, 1988, Huang and Wang, 2004), point to the fact that searchers assess the relevance of documents relative to the relevance of documents they have already seen, Spink et al. showed that relevance judgments may vary according to the assessor's developing understanding of relevance for a topic (Spink et al., 1998) and Voorhees showed that agreement between assessors demonstrated considerable variation, (Voorhees, 2000).

With TREC data we cannot assess how subjective factors affect the *process* of searching as there is no interactive searching within most tracks, only assessment of what has been submitted for assessment. However we can examine how subjective factors affect relevance assessment.

We focus in this case on the HARD (High Accuracy Retrieval of Documents) track which investigated how contextual information could improve retrieval performance (Allan, 2003; Allan, 2004). In section 3.1.3 we discuss how we might use data from other tracks. In this track, each participating group could submit a clarification form, an html form containing questions to be answered by the assessor, to gain additional information from the assessor which may be useful in improving an initial baseline retrieval run.

The TREC assessors had no more than 3 minutes on each form so this was an interesting test of how much useful information could be gained from an assessor in a short period of time and what types of information were useful in improving retrieval performance. As explained in the TREC overview reports, the use of these forms showed that even this limited form of interaction could generally increase retrieval effectiveness although always not to the level of good automatic techniques and also that different queries may be served better by different approaches (Allan, 2003; Allan, 2004).

For the analyses presented in sections 3.1.1 and 3.1.2 we perform a series of Canonical Correlation Analyses (CCAs), multivariate tests for understanding the relationship between multiple variables in a controlled experimental study (Sherry and Henson, 2005). A CCA investigates the relationships between variables where each variable may have multiple causes (e.g. low search precision may arise for multiple reasons) and multiple effects (e.g. high interest in a topic may lead to different search behaviour *and* to the use of different relevance criteria than when the assessor has low interest in a topic).

At its most simple level a CCA is a Pearson *r* correlation between two synthetic, or latent, variables both of which represent a set of variables. A particular advantage of CCAs is that

they reduce the likelihood of so-called Type 1 errors where a significance tests shows a statistically significant result where there is no significant relationship between variables (Sherry and Henson, 2005). This is a particular risk when running multiple significance tests over the same data. CCAs run the tests simultaneously reducing, although not eliminating, the likelihood of such errors.

In section 3.1.1 we examine contextual information provided by the assessors in response to direct questions from participating groups and in section 3.1.2 we examine contextual factors given in the topic narrative.

3.1.1. Hard 2005

Our first study examines the impact of assessor's personal relationship to the topic and the relevance assessments provided for that topic. That is, do subjective factors, such as topical knowledge, affect the type of assessments made by the assessors on the documents being assessed? In terms of a CCA analysis this is an investigation between the relationship between two sets of variables: one reflecting the assessor's personal relationship to the topic and one reflecting the relevance assessments made.

The first set of variables come from the assessor's responses to selected questions from the clarification forms, specifically the responses to the HARD-2005 STRA3 and UWAT1 clarification forms which asked about the assessor's **confidence**[1] in assessing relevance, the assessor's **interest** in the topic, the assessor's level of **specific** knowledge on the topic and the assessor's level of **familiarity** on the topic. That last two questions were framed differently and there was a low correlation between the responses to the questions (r=0.092, Pearson) so it seems fair to treat them as separate items. These factors represent contextual information about the assessors and their relationship to the topic.

The second set of variables reflects three post-search outcomes:

- the overall **precision** measured as the number of documents assessed as being either relevant or highly relevant divided by the total number of documents assessed for each topic. Precision is investigated as being one of the main search outcomes from any user or system study.
- the proportion of documents that were assessed as relevant (as opposed to highly relevant) which will be referred to as **percentpartial**. Although the term partial relevance was not used in this track we use this label for relevant documents to differentiate them from highly relevant documents. This measure is included due to previous evidence, (Spink et al., 1998), that differences between the proportion of highly relevant documents and partially relevant documents during a search can highlight meaningful differences in how searchers have defined relevance. This is discussed in more detail below.
- The number of query **expansion** terms selected from the clarification forms. This is described more fully in section 3.2. The selection of expansion terms here is treated as an outcome variable as, like the other outcome variables, it is an observed variable rather than a subjective variable.

[1] Ruthven et al. (Ruthven et al., 2007) suggested that confidence should be treated as a binary variable measuring the assessor's willingness to declare their confidence rather than their actual level of confidence in performing relevance assessment for a topic. We also ran a CCA with confidence coded as a binary variable reflecting confident/not confident. However the results did not differ substantially from what is reported in Table 1.

The CCA technique produces functions (or models) to maximise this correlation by weighting the variables to highly weight those variables that contribute most to the correlation. CCAs may produce multiple models and usually produce one model for each dependent variable. Naturally, we are only interested in statistically significant models, ones which describe a significant relationship between the two latent variables.

The analysis produced one significant CCA, Table 1. The squared canonical correlation R_c^2 for the function was 0.525, F(12, 106.12)=2.07, p=0.002, the full model explaining a substantial portion, about 48%, of the variance shared between the variable sets.

Table 1 has four columns: the dependent (assessor) and independent (outcomes) variables; *coef* which are the standardized canonical functional coefficients, weights derived during the process of maximising correlations and which are analogous to weights used in regression analysis; structure coefficients (r_s) which are correlation weights between the observed variable and synthetic variable and can be seen as a measure of how much the observed variable contributes to the creation of the synthetic variable; finally the communality coefficient (h^2) is a measure of the variance explained by the variable, essentially how useful the variable was in creating the correlation between latent variables. Following convention, variables whose structure coefficients or communality coefficients are above 0.45 are highlighted as these contribute most to the model.

	coef	r_s	h^2 (%)
Assessor variables			
STRA3 confidence	-0.004	-0.105	1.10%
STRA3 interest	-0.431	**-0.668**	44.60%
STRA3 specific	-0.602	**-0.755**	**56.94%**
UWAT1 familiarity	0.560	**-0.456**	20.83%
Outcome variables			
precision	-0.671	**-0.722**	**52.05%**
percentpartial	0.634	**0.676**	**45.67%**
expansion	0.433	0.203	41.14%

Table 1: CCA for Hard 2005

From Table 1 we can see that the strongest predictor variable was the participants' declared specific knowledge on the topic before assessing any returned documents. This has strong structure coefficient and function coefficients meaning that the variable has a large impact in characterising the pre-search synthetic variable and creating the correlation between the assessor and outcome variables. The level of interest has a high structure coefficient which indicates that it is important in explaining much of the variance in the model. Familiarity has a moderate function coefficient and structure coefficients whereas the coefficients for confidence are extremely low.

The result for interest, a modest function coefficient but large structure coefficient, suggests some relationship with the other variables, i.e. that interest and familiarity with a topic are somehow related but not identical. A follow-up correlation reveals a significant but not strong

correlation between these two variables (r=0.467, p<0.001, Spearman's Rho). This provides an interesting side-question of the relation between these two variables when creating and analysing search topics.

The strongest contributors to the outcome synthetic variable were the overall **precision** and **percentpartial**. The low contribution of **expansion** indicates that the level of interactivity, in this case as measured by choosing expansion terms, does not contribute much to the outcomes synthetic variable and does not relate strongly to the assessor characteristics.

The signs for **familiarity**, **interest** and **specific** are similar to **precision** but opposite to that of **percentmarginal**. Taken together the results indicate that topics where the assessor has low levels of interest, familiarity and specific knowledge are those that are likely to have low precision and a high use of partial relevance assessments, i.e. low use of the highly relevant category. Conversely, those topics where the assessor is more familiar with the topic, are more interested in the topic and have some topical knowledge are those that are likely to have the highest precision after assessments – the assessor marking more of the pool as relevant - and higher use of the highly relevant category. This demonstrates how changes in assessor profile can change the nature of relevance assessments given during the TREC assessment process and thus impact on evaluation itself by changing the nature of the relevant documents used for evaluation.

This finding also matches the findings of Spink et al. who showed that a high number of assessments of partial relevance indicate situations where assessors are unsure of precisely how to define relevance and who need to work more closely with the retrieved documents to determine how they should assign relevance categories (Spink et al., 1998). It may also go some way to explaining the results obtained by Scholer et al. (Scholer et al., 2011) on inconsistency in relevance assessments; inconsistency may be the result of a learning process, learning for a topic how relevance is to be defined, rather than a simple matter of error.

3.1.2. Hard 2003

Our second study examines the impact of assessor's pre-stated requirements for relevance on the relevance assessments provided for that topic. That is, when assessors require a precise level of specificity in a response, does this affect their assessment behaviour?

To investigate this, we use data from the 2003 HARD track in which topics contained meta-data about the assessor's requirements for relevance (Allan, 2003). We focus here on the

- **purpose** variable which represents why the assessor is looking for information: a value of *any* means that the assessor has no particular purpose in mind, a value of *background, details* or *answer* indicates the assessor is for a particular type of answer.
- **genre** variable which represents the type of material the assessor is interested in: values of *overview, reaction i-reaction* or *administrative* mean the assessor is interested in specific types of material whereas a value of *any* indicates that any genre is acceptable or none was indicated.
- **granularity** variable which captures the amount of text that the searcher is anticipating in a response with values *document, passage, sentence* or *phrase* meaning that the assessors would like specific units of text and a value of *any* means the user has no specific granularity in mind or did not specify one.

In HARD 2003 assessors were asked to differentiate between *hard* and *soft* relevance: hard relevant documents were relevant and satisfied these meta-data requirements whereas soft relevant documents were relevant but did not satisfy the meta-data requirements. Additional meta-data information, such as familiarity, was not used in the analysis presented here due to the low variability in these variables making them unsuitable for a CCA.

Creating a CCA between **purpose**, **granularity** and **genre**, as answer variables, and **precision** (the number of hard and soft relevant documents divided by the total number of assessed documents) and **percentsoft** (the proportion of the relevant documents that are soft relevant) created only one very weak model.

A second CCA was constructed by recoding the granularity and genre variables to reflect topics for which values for these requirements were not specified (values of *any*) or were specified (any other value) in the topic description. This separates out situations where the assessor had a good idea of what responses were required before assessing the responses and situations where the assessor was more flexible in relevance definitions. There were no topics with a value of *any* for the *purpose* variable. This produced one significant model, Table 2. The squared canonical correlation (R_c^2) for the function was 0.21, F(6,88.00)=3.29, p=0.006, the full model explaining about 33% of the variance shared between the variable sets.

	coef	r_s	h^2 (%)
Answer variables			
purpose	-0.848	**-0.740**	**54.76%**
genre	0.354	0.178	3.17%
granularity	-0.595	**-0.521**	27.14%
Outcome variables			
precision	0.603	**0.582**	33.87%
percentsoft	0.813	**0.798**	**63.68%**

Table 2: CCA for Hard 2003

The strongest contributions to the answer variables was the **purpose** which had a strong function coefficient and structure coefficient and **granularity** which made a smaller, though important, contribution. The strongest contribution to the outcome variable was **percentsoft**. What this indicates is that, for topics where the assessors have more specific criteria for relevance, then both precision and the proportion of 'soft' relevant documents are low. On the other hand, for topics where the assessors have fewer pre-defined meta-data values both the precision and use of soft relevance are high. The first finding makes sense in terms of the track description: where assessors have stricter relevance criteria then we should expect fewer documents to match these criteria and also to see fewer documents that do *not* match these criteria to be judged relevant. Similarly when assessors are less strict in their prior expectations of what a good answer will look like, they are more open to employing more liberal relevance criteria. This also matches findings from previous user studies, such as that of Saracevic and Kantor (1988) who found that more specific tasks resulted in lower precision whilst more general tasks result in higher precision.

3.1.3. Summary

Most TREC studies evaluate query and systems attributes and their impact on system evaluation; here we showed that the assessors are also important. Similarly there are

differences in the type of assessments (e.g. making judgments of partial relevance vs. high relevance) made under different situations and how we use these assessments could affect our evaluation of the systems being investigated. Such questions often form part of user study research questions, here we show they are valid in TREC as well.

In section 2 we mentioned Saracevic''s manifestations of relevance (Saracevic, 1996). Most system evaluations operate at the lower relevance levels described by Saracevic matching the content of documents against a query or query representation, such as a narrative in the case of TREC. Users of IR systems often operate at higher relevance levels which take into account aspects such as novelty, accessibility and interest. An important implication of Saracevic's proposal is that users may be using different criteria to judge the quality of a system's response than those used in system evaluations: test collections tell us how well systems retrieve topically relevant documents whereas we evaluate systems according to how well they retrieve documents that are of use and of interest. Studies such as the ones here could tell us more about how these contextual factors may arise in TREC and give a more useful interpretation of system performance currently offered by test collection analyses.

The meta-data we used here came directly from the assessors themselves. However, we may also *assign* attributes to topics, characterising them by their complexity or specificity, to allow analyses of how pre-determined characteristics affect such relevance decisions. This would allow investigations of the kind described by Sormumen into the use of stricter and liberal relevance thresholds in TREC assessments (Sormunen, 2002).

The fact that some of these results match those found in user studies may also mean that the criticism that TREC is unrealistic, because of the way it collects relevance assessments, may not be valid and that TREC assessors are behaving in similar ways when assessing relevance to assessors in more naturalistic studies, albeit in a more controlled setting.

3.2. HARD interactive query expansion

Our second study looks at interactive query expansion, a perennial topic of interest within IR and one that has figured in many different TREC tracks, including the HARD, ciQA and interactive tracks, giving a range of data that may be used to explore query expansion in different controlled settings. As an example of the kind of study that may be possible with TREC data, we use data from the TREC 2005 HARD track to investigate intra-assessor reliability in making query expansion decisions. Like Lin and Zhang, (2007) we examine intra-assessor consistency rather than inter-assessor consistency to examine how often an assessor makes the same decision multiple times.

In TREC-2005 several clarification forms asked the assessors to select, or rate, expansion terms from a supplied list. Many expansion terms appeared on several clarification forms and the repeated assessment of expansion terms by the same assessors on the same topic allow a comparison of how consistent is this term selection. So our research question is: are the same expansions terms consistently selected or is selection of expansion terms more random? If expansion term selection is highly consistent then this would provide evidence for the importance of expansion terms in describing what information the assessor wants returned; if expansion term selection is not consistent then we may need to reinterpret how useful such information actually is in deciding what information should be retrieved and perhaps treat expansion terms differently from user-generated query terms.

To test this, for each topic, we took all the offered terms from clarification forms supplied as part of the CASP1, CASP2, CASS1, INDI1, NCAR1, NCAR2, UIUC1, UIUC2, UIUC3, UWAT2 submissions. We then pooled the terms and noted how often each term was selected as good term by the assessors. That is, on how many forms was the term selected as being a good term and how often was it not selected as useful?

Table 3 shows the basic breakdown of how many terms were offered on multiple forms. The **times offered** column indicates on how many forms a term appeared (on only form, on two forms, etc.); the **number of terms** indicates how many terms were offered multiple times (4922 terms appeared on only form, 1309 terms appeared on two forms, etc.); **at least one accepted** indicates how many times at least one of these offerings were accepted as useful by the assessor (e.g. of the 4922 terms that only appeared on one form, 688 were accepted as useful, of the 1309 terms that appear on only two forms there were 415 terms which were selected as useful on least one form, etc.). The percentage of accepted terms to offered terms gives the **acceptance rate**.

Times offered	Number of terms	At least one accepted	Acceptance rate
1	4922	688	13.98%
2	1309	415	31.70%
3	1156	414	35.81%
4	363	232	63.91%
5	160	113	70.63%
6	113	96	84.96%
7	87	81	93.10%
8	65	62	95.38%
9	30	29	96.67%
10	7	7	100.00%
11	7	6	85.71%
12	2	2	100.00%
13	4	3	75.00%

Table 3: Term offerings and acceptance rates

Not surprisingly, terms that are offered on more forms are more likely to have at least one of these offerings accepted as being useful and there is a strong, significant correlation (r=0.787, $p<0.01$, Spearman's Rho) between the number of times a term was offered and the acceptance rate. The relationship between the acceptance rate and the number of times offered is roughly linear, e.g. terms that are offered 4 times are approximately 4 times as likely to be accepted (at least once) as useful compared to terms that are only offered once. The relationship between **at least one accepted** and **acceptance rate** approximates a power law distribution.

In Table 4 we focus on those terms that are offered multiple times and where at least one term offering was selected as being useful. The **times offered** and **number of terms** columns are as in Table 3. The **average** column indicates the average selection rate: terms that were offered on three forms, on average, would be selected as good on two out of the three forms (65.86%). The final column, SD (standard deviation), indicates how variable is this average.

As can be seen in the average column there is a similar distribution irrespective of how many times a term is offered: if a term is accepted as useful, approximately 75% of the time it is offered it will be accepted and the remaining 25% of times it will not be accepted. For the conditions where there is sufficient data to generalise (up to 9 times offered) the consistency

of term selection is very steady regardless of how many times the term is offered. This suggests a certain error rate in being consistent. So although being offered more times increases the likelihood of a term being accepted at least once, if a term is accepted then the rate at which it is accepted is independent of how often it is offered. The intra-assessor consistency rate of 75% is sufficiently high that we can assert that the assessors are consistently selecting similar terms for expansion on each form.

Times offered	Number of terms	Average	SD
2	415	73.37%	24.98%
3	414	65.86%	27.14%
4	232	70.47%	27.54%
5	113	74.34%	27.71%
6	96	74.13%	28.90%
7	81	77.60%	25.95%
8	62	77.62%	24.81%
9	29	80.84%	21.80%
10	7	88.57%	10.69%
11	6	89.39%	6.84%
12	2	66.67%	47.14%
13	3	84.62%	7.69%

Table 4: Term offerings and selection of accepted terms

In Table 5 we show the average for each assessor. On the whole, assessors do mark the majority of terms offered multiple times as being relevant. However, a Kruskal-Wallis test showed a significantly different ($p<0.001$) rate of acceptance *between* the assessors indicating that some assessors are more consistent than others in which terms they see as important. Assessor 5, for example, is far less consistent in term selection than assessor 6 suggesting there may be factors worth investigating in why some people are more consistent than others in query expansion decisions.

	Assessor					
Times offered	**1**	**2**	**3**	**4**	**5**	**6**
2	76%	64%	75%	76%	74%	80%
3	64%	63%	71%	66%	59%	73%
4	75%	77%	70%	72%	65%	77%
5	66%	88%	72%	68%	61%	78%
6	60%	83%	80%	82%	68%	80%
7	59%	79%	91%	85%	58%	87%
8	61%	93%	84%	96%	71%	79%
9	67%	76%	86%	79%	33%	93%
10	-	96%	-	90%	80%	-
11	-	91%	91%	-	36%	91%
12	-	100%	-	-	-	-
13	-	85%	-	92%	-	-

Table 5: Selection rates by assessor

These results aggregate over a large number of term offerings. In order to detect whether some *types* of terms are selected more consistently we investigated several term characteristics. Firstly, we considered whether terms used in the topic description would be

selected more consistently. As demonstrated in several user studies, e.g. (Fang, 2001), terms used in written search topics are often used by experimental participants in creating and modifying queries. To test whether there was a preference to select terms that appeared in the topic description we ran the same analysis as in Table 4 but only for terms which were selected as being useful **and** which appeared in the topic description, Table 6.

On average, from Table 4, if a term is accepted as being useful then approximately 75% of offerings are selected as useful. However, if the term appears in the topic description supplied to the assessor then this average rises to approximately 85%, Table 6, so a term that appears in the topic description is more likely to be selected as useful for query expansion. This difference in acceptance rate is statistically significant ($p<0.05$, Wilcoxon Signed Rank) and the standard deviation for acceptance of terms in the topic description is significantly lower than for terms not in the topic description ($p<0.05$, Wilcoxon Signed Rank).

These two results indicate that the assessors are perhaps primed towards seeing some terms are more useful by their presence in the topic description and that they are more consistent in their selection when these terms appear in multiple clarification forms. In this track, assessors were generally not assessing their own topics (although they were allowed to interpret relevance for themselves) so perhaps, as in user studies, the assessors were relying more heavily on the topic description than their own personal knowledge when making expansion term decisions.

Times offered	Number of terms	Average	SD
2	37	85.14%	23.17%
3	43	75.97%	25.54%
4	40	76.25%	24.64%
5	30	88.00%	16.27%
6	15	91.11%	13.90%
7	27	87.83%	11.69%
8	21	80.95%	24.56%
9	30	80.18%	25.10%
10	4	90.00%	8.16%
11	3	93.94%	5.25%
12	1	100.00%	-

Table 6: Acceptance rates for terms in the topic description

Secondly we investigated whether more common terms - those that are more common in the document collection - are more likely to be assessed consistently. Common terms are usually more recognizable and therefore it may be easier to judge whether they relate to the topic of not; conversely terms that are less common, such as names, may be seen as unusual, display a less clear connection with the topic and therefore may be more difficult for the assessor to predict their value.

We tested this in two ways: firstly a straight correlation test between the rate of acceptance of terms and their collection occurrence which showed almost no correction (r=-0.028, p=0.135, Spearman's Rho). Secondly we split the terms in groups, as shown in Table 7 where we have split the terms into quartiles according to their occurrence within the document collection (1-25% are the least frequent quartile of terms, 26-50% are the second least frequently occurring

terms, etc.). The **average acceptance** column gives the average acceptance rate of the terms in each quartile. As can be seen from Table 7 there is no obvious relationship between how often a term appears in a collection and how often that term is accepted as being useful on a clarification form. Neither is there any significant difference in acceptance rate between the occurrence categories (p=0.816, Kruskal-Wallis). So terms that are more frequent are not more likely to be consistently selected for query expansion.

Collection occurrence	Acceptance rate	SD
1-25	71.71	26.71
26-50	71.18	26.55
51-75	72.06	26.04
76-100	71.07	26.69

Table 7: Acceptance rates for terms in the topic description

This case study examined relevance behaviour concerning how TREC assessors would behave towards terms that were offered multiple times on the same topic. As noted by Allan (Allan, 2005) this is not an interactive information retrieval study in the sense that searchers were interacting with possible expansion terms and then seeing the results of expansion decisions in the form of which new documents were retrieved. However, it is a relatively large study, compared to most user studies, of assessors making expansion term decisions.

The repeated judgments of the same terms have provided new findings which can be summarised as follow:

- Offering an expansion term more often increases the chance of at least once occurrence of that term being selected as useful. This is not surprising in itself and the absence of such a finding may invalidate the use of clarification forms as it would suggest the assessors were not treating each form as a new series of judgments.

- If at least one occurrence of a term is selected as being useful then the same *proportion* of occurrences will be selected, independently of how often the term is offered. What this suggests is that there is a baseline rate of consistency in selecting the same term as being useful multiple times. That is, a consistent rate of intra-assessor variation in selecting useful expansion terms.

- Terms which appear in the topic description, and are offered multiple times, are more likely to be selected as useful. This result has appeared in interactive studies before and has raised questions about how to provide search tasks to participants without biasing them towards some terms over others. That TREC topics are themselves a useful source of expansion terms is not surprising given the nature of the topic descriptions however it poses the question of how we separate out this 'description effect', i.e. that the experimental protocol itself may interfere with behaviour, both in interactive studies and analysis of TREC data.

- The likelihood of a term being accepted multiple times does not appear to be related to how often the term appears in the collection but is related to, as yet unknown characteristics of, the person who is making the assessment.

The HARD and later ciQA tracks provide a particularly rich source of data on relevance behaviour in TREC as they provided a means to obtain information directly from the assessor. This includes subjective information such as their familiarity with the topic, as in section 3.1.1 and additional topic-related information such as the provision of additional free text search terms. This information can be used to investigate many research questions about relevance behaviour that could guide future user studies.

3.3. Document factors and assessment

The third case study looks at the document factors and their possible effect on relevance assessment.

In this study we examined the following factors:

1. the total number of documents in the assessment pool for each topic (**pool size**)
2. the average length of the documents in the assessment pool for each topic (**length**)
3. the **assessor** who assessed the topic using the codes provided by TREC. This variable encapsulates the various factors identified in section 3.1 into a single factor.
4. the **emotional density** of the relevant and non-relevant documents as explained in section 3.3.1.

3.3.1. Emotional density

Barry and Schamber in their landmark studies of relevance assessment outlined around 80 relevance criteria: reasons why a searcher may mark a document as relevant (Barry and Schamber, 1998). One of these criteria is the affective response of the assessor. In their studies the affective response, the degree to which the reader emotionally responds to a document, could lead to a document being judged relevant or not relevant. Such criteria can form part of the higher relevance levels described by Saracevic where the motivations and preferences of the user can affect relevance decisions (Saracevic, 1996).

Affective responses may come into play in situations where searchers have the freedom to make relevance decisions that are not solely based on topical decisions, such as in leisure searching. In a TREC context where the topic narratives give detailed descriptions of the content that makes a document relevant or non-relevant the emotional content of documents may not be a factor in determining relevance.

However there is evidence from fields such as consumer research, e.g. (Lau-Gesk and Meyers-Levy, 2009) that affective cues within texts can affect behaviour including evaluation judgments. There is also a long-standing debate within psychology regarding the relationship between cognitive and emotional attention demonstrating that negative emotional cues are more difficult to ignore than positive ones and that cognitive performance may change in the presence of certain emotional, but non-task related, cues, (Compton, 2003). This suggests that the emotional language of documents may lead to differing decisions on relevance.

To investigate this, we used the The Compass DeRose Guide to Emotion Words[2], a list of 800 emotional words, as a starting point to provide a list of emotional words. From this list we removed words for which the primary meaning of the word was not emotional, e.g. 'blue', and added variants of existing words such as stemmed variants to provide a list of 757 emotional words. Using this emotional word list we then counted the emotional density of the TREC topics: the ratio of emotional words to non-emotional words in the relevant and non-

[2] http://www.derose.net/steve/resources/emotionwords/ ewords.html

relevant documents for each topic as a measure of how emotionally dense were the relevant and non-relevant documents on average.

3.3.2. Analysis

The data in this case study comes from the TREC HARD 2005 track. Our analysis is between what we refer to as input variables which represent the inputs to the TREC assessment process, namely the documents, topics and assessors, and the outcome from the assessment process, namely the precision and percentage of relevant documents that are assessed as partially relevant.

This produced one significant model, Table 8. The squared canonical correlation (R_c^2) for the function was 0.565, F(10,76.00)=2.51, p=0.012, the full model explaining about 44% of the variance shared between the variable sets.

The strongest predictor variable was the **assessor** variable. This has strong structure coefficient and function coefficients meaning that the variable has a large impact in characterising the pre-search synthetic variable. Both the average **length** of the assessed documents and the **emotional density** of the relevant documents had moderate function coefficient and structure coefficients whereas the coefficients for the emotional density of the non-relevant documents, topic difficulty and pool size are extremely low.

The strongest contributors to the outcome synthetic variable were the percentage of documents assessed as being **partially relevant** with precision making a moderate contribution.

This indicates that the assessor is the main variable in the use of partial relevance assessments. This also fits with the findings of Ruthven et al., (2007) who found that differences in the assessor's *relationship* towards the topic correlates with the use of partial relevance assessments. The coefficients for length and emotional density of relevant documents suggests some relationship to the outcome variables; namely that for some assessors, assessing long, emotionally sparse documents (i.e. low emotional density in relevant documents) is associated to lower precision and lower use of the high relevance category. Or, in other words, long documents with little emotional content may be associated less likely to initiate a decision of high relevance.

In Barry's study of relevance criteria, affective responses were one of the most numerous criteria used to judge relevance. The affective nature of a document and the affective response of an assessor are not necessarily related, however this relationship would seem worthy of further research.

	coef	r_s	h^2 (%)
input variables			
pool size	-0.140	-0.127	1.60%
length	-0.444	**-0.541**	29.23%
emotional density relevant documents	-0.279	**-0.515**	26.53%
emotional density non-relevant documents	-0.050	-0.272	7.40%
assessor	0.758	**0.772**	**59.61%**
outcome variables			
precision	0.577	**0.470**	22.10%

percent partially relevant	0.889	**0.820**	**67.16%**

Table 8: CCA for Hard 2005

4. Discussion

As Saracevic noted in his 2007 article, '*Relevance is timeless. Concerns about relevance will always be timely*' (2007a). Information seeking and information retrieval have taken divergent interests in relevance. Information Retrieval (IR), with its perennial concerns over evaluation and evaluation methodology, has concentrated on what Saracevic referred to as '*the effects of relevance*' (Saracevic, 2007a) or how relevance, in the form of relevance judgments can impact on the evaluation of retrieval systems. Information seeking has tended to focus on the '*the behavior of relevance*' (Saracevic, 2007a) to better understand the human process of making relevance judgments and the factors that may affect these judgments.

The major approach we take in IR is to create research hypotheses and then test them, usually by creating new user tests or testing the hypotheses with data from initiatives such as TREC. A complementary research agenda could be to ask: what relevance questions can be answered by existing TREC data and what relevance behaviours are already contained within TREC but remain to be discovered? A particular use of TREC might be to perform systematic meta-analyses: asking critical questions of relevance and examining to what degree the TREC assessments and studies help answer these questions.

The advantages of the TREC data are not to be ignored: it is large scale, it is gathered in controlled settings, topics are often re-used which would allow comparison between different people assessing the same topics, individual assessors often participate in multiple TRECs thereby creating a lot of data in different assessment situations and the data is freely available.

There are of course objections to such a research agenda:

- As noted in section 2 TREC is often criticised for being unrealistic with respect to end-user searching because of the fixed relevance narratives. We have shown, for example in section 3.1., that even though attributes of relevance are predefined, assessor characteristics in TREC are important as they have been shown to be in user studies.
- TREC focuses on assessment rather than interactive searching. However, this is true of many important studies of relevance behaviour, e.g. (Barry and Schamber, 1998). The concentration of TREC on assessments, abstracting away, in most tracks, from interaction offers the possibility to focus on relevance assessment. Many experiments in areas such as psychology use very controlled study designs to learn about phenomena before investigating the phenomena in contextualised studies. There is no reason why we cannot do the same here to exploit the wealth of quantitative data on relevance before conducting smaller-scale, qualitative user studies.
- TREC assessors are homogenous with high levels of experience in information assessment and training in assessment. This is also true of many information seeking and retrieval studies which use university students as the main participant group. It is also the case that such studies are often 'artificial' in using pre-defined search tasks. As with any study that involves a particular demographic we need to be careful about what generalisations we can make and which we cannot. We are desperately short of validation studies using different demographics within information seeking and

retrieval. This, however, does not mean we cannot learn about relevance and relevance behaviour from TREC.

The case studies presented here are only a small sample of those that are possible with TREC data. Not all research questions can be answered and the study designs of individual tracks may limit which relevance questions can be answered and which cannot. We also have to be careful of the context of the tracks themselves which have specific aims and designs. Here we have investigated a small sample of the available data and acknowledge that conclusions from one data set may not apply to other data sets because of the track's detailed design.

However, the sheer variety of tracks provides a rich seam of data for some questions, especially if we can aggregate across tracks. Other fields, particularly in areas such as social science and biomedical sciences see the re-use of data and secondary data analysis as sign of maturity of a discipline. Given how much effort we have placed in TREC and how much the assessors have contributed to producing relevance assessments it is perhaps worth making full value of this investment by learning as much as we can from TREC.

Two complementary criticism have become common: information seeking studies are too small scale to have any real predictive power in explaining relevance (the systems perspective) and initiatives such as TREC are too artificial to have any external validity in explaining relevance (the information seeking perspective), with Karen Spärck Jones providing a good commentary of realism within TREC in (Spärck Jones, 1996). In his 1999 article Nigel Ford tackled this dichotomy head-on:

'…, *much research in information science has arguably provided highly reliable answers to highly meaningless questions. ...Without critical interaction with complementary perspectives the increasing use of subjective analysis of introspections using small samples of information users threatens to supply highly meaningful questions with highly unreliable answers. Some balance and integration must be achieved.*' Ford (1999, p1151)

In this article we hope to provide some steps towards this '*balance and integration*' by proposing that the investment from the IR community in the form of TREC (and other initiatives such as INEX and FIRE) can be used creatively to answer questions on relevance by the information seeking community and to better understand the results of system evaluations by the information retrieval community.

5. References

Allan. J. (2003), "Hard track overview in trec 2003: High accuracy retrieval from documents", In Voorhees, E. M. and Buckland, L. P. (Eds), *Proceedings of the Twelfth Text REtrieval Conference, TREC 2003,* Special Publication 500-255, National Institute of Standards and Technology (NIST), pp. 24-37.

Allan. J. (2004), "Hard track overview in trec 2004 – high accuracy retrieval from documents", In Voorhees, E. M. and Buckland, L. P. (Eds), *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*, Special Publication 500-261, National Institute of Standards and Technology (NIST), pp. 25-35.

Allan. J. (2005), "Hard track overview in trec 2005 high accuracy retrieval from documents", In Voorhees, E. M. and Buckland, L. P. (Eds), *Proceedings of the Fourteenth Text REtrieval*

Conference, TREC 2005, Special Publication 500-266, National Institute of Standards and Technology (NIST), pp. 51-68.

Barry, C. L. and Schamber, L. (1998), "Users' criteria for relevance evaluation: a cross-situational comparison", *Information Processing & Management*, Vol. 34 No. 2-3, pp. 219-236.

Bateman, J. (1998), "Changes in relevance criteria: A longitudinal study". In Preston, C. M. (Ed.), *Proceedings of the ASIS annual meeting*, volume 35, pp. 23-32.

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

Chu, H. (2007), "Factors affecting relevance judgment: a report from the trec legal track", *Journal of Documentation*, Vol. 6 No. 2, pp. 264-278.

Compton, R., M. Banich, A. Mohanty, M. Milham, J. Herrington, G. Miller, P. Scalf, A. Webb, and Heller, W. (2003), "Paying attention to emotion: an fmri investigation of cognitive and emotional stroop tasks", *Cognitive*, Affective, & *Behavioral Neuroscience*, Vol. 3 No. 2, pp. 81-96.

Eisenberg, M. and Barry, C. (1988), "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.

Fang, X. (2001), "Web searching behavior: Selection of search terms", In Cheok, A. D. And Chittaro, L. (Eds), *Proceedings of the Ninth International Conference on Human-Computer Interaction*, Vol 1, pp. 898-902.

Florance, V. and Marchionini, G. (1995), "Information processing in the context of medical care", In Fox, E. A., Ingwersen, P. And Fidel, R. (Eds), *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 158-163.

Ford, N. (1999), "The growth of understanding in information science: Towards a developmental model", *Journal of the American Society for Information Science*, Vol. 50 No. 12, pp. 1141-1152.

Ford, N., Miller, D., and Moss, N. (2001), "The role of individual differences in Internet searching: An empirical study", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 12, pp. 1049-1066.

Gwizdka, J. and Lopatovska, I. (2009), "The role of subjective factors in the information search process", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 12, pp. 2452-2464.

Harter, S.P. (1996), "Variations in relevance assessments and the measurement of retrieval effectiveness", *Journal of the American Society for Information Science*, Vol. 47 No. 1, pp 37-49

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. 970-979.

Kelly, D., Fu, X., and Shah, C. (2010), "Effects of position and number of relevant documents retrieved on users' evaluations of system performance", *ACM Transactions on Information Systems*, Vol. 28 No. 2, pp. 1-29.

Lau-Gesk, L. and Meyers-Levy, J. (2009), "Emotional persuasion: When the valence versus the resource demands of emotions influence consumers' attitudes", *Journal of Consumer Research*, Vol. 36 No. 4, pp. 585-599.

Lin, J. and Zhang, P. (2007), "Deconstructing nuggets: the stability and reliability of complex question answering evaluation", In Clarke, C. L. A., Fuhr, N., Kando, N., Kraaijj, W., and de Vries, A. P. (Eds), *Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval,* pp. 327-334.

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

Ruthven, I., Baillie, M., and Elsweiler, D. (2007). "The relative effects of knowledge, interest and confidence in assessing relevance", *Journal of Documentation*, Vol. 63 No. 4, pp. 482-504.

Saracevic, T. (1996), "Relevance reconsidered. Information science: Integration in perspectives", In Ingwersen, P. And Pors, N. O. (Eds), *Proceedings of the Second Conference on Conceptions of Library and Information Science*, pp. 201-218.

Saracevic, T. (2007a), "Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestation of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 1915-1933.

Saracevic, T. (2007b), "Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 126-2144

Saracevic, T. and Kantor, P. (1988), "A study of information seeking and retrieving. II. Users, questions, and effectiveness", *Journal of the American Society for Information Science*, Vol. 39 No. 3, pp. 1097-4571

Scholer, F., Turpin, A. and Sanderson, M. (2011), "Quantifying test collection quality based on the consistency of relevance judgements", In Ma, W.-Y., Nie, J.-Y., Baeza-Yates, R., Chua, T.-S., and Croft, W. B. (Eds), *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 1063-1072.

Sherry, A. and Henson, R. K. (2005), "Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer", *Journal of Personality Assessment*, Vol. 84 No. 1, pp. 37-48.

Smith, C. L. and Kantor, P. B. (2008), "User adaptation: good results from poor systems", In Myaeng, S. H., Oard, D. W., Sebastiani, F., Chau, T.-S., and Leong, M.-K., (Eds), *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 147-154.

Smucker, M. D. and Jethani, C. P. (2010), "Human performance and retrieval precision revisited", In Chen, H.-H., Efthimiadis, E. N., Savoy, J., Crestani, F., and Marchand-Mallet, (Eds), *Proceedings of the 33rd international ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 595-602.

Smucker, M. D. and Jethani, C. P. (2011), "Measuring assessor accuracy: a comparison of nist assessors and user study participants", In Ma, W.-Y., Nie, J.-Y., Baeza-Yates, R., Chua, T.-S., and Croft, W. B. (Eds), *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1231-1232.

Spärck Jones, K. (2006), "What's the value of TREC: is there a gap to jump or a chasm to bridge?", *SIGIR Forum*, Vol. 40 No. 1, pp. 10-20.

Sormunen, E. (2002), "Liberal relevance criteria of trec -: counting on negligible documents?", In Järvelin, K., Beaulieu, M., Greisdorf, H., and Bateman, J. (Eds), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 324-330.

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

Turpin, A. H. and Hersh, W. (2001), "Why batch and user evaluations do not give the same results",In Kraft, D. H., Croft, W. B., Harper, D. J., and Zobel, J. (Eds), *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 225-231.

Vakkari, P. and Hakala, N. (2000), "Changes in relevance criteria and problem stages in task performance". *Journal of Documentation*, Vol. 56 No.5, pp. 540-562.

Voorhees, E. M. (2000), "Variations in relevance judgments and the measurement of retrieval effectiveness". *Information Processing & Management*, Vol. 36 No. 5, pp. 697-716.

Voorhees, E. M. (2001), "Evaluation by highly relevant documents", Kraft, D., H., Croft, W. B., Harper, D. J. and Zobel, J. (Eds), *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 74-82.

Voorhees, E. M. and Harman, D. K. (2005), *TREC: Experiment and Evaluation in Information Retrieval*, The MIT Press. Cambridge, MA.