

# Liberal Relevance Criteria of TREC – Counting on Negligible Documents?

Eero Sormunen  
University of Tampere  
FIN 33014 University of Tampere  
Finland  
+358-3-2156972  
Eero.Sormunen@uta.fi

## ABSTRACT

Most test collections (like TREC and CLEF) for experimental research in information retrieval apply binary relevance assessments. This paper introduces a four-point relevance scale and reports the findings of a project in which TREC-7 and TREC-8 document pools on 38 topics were reassessed. The goal of the reassessment was to build a subcollection of TREC for experiments on highly relevant documents and to learn about the assessment process as well as the characteristics of a multigraded relevance corpus.

Relevance criteria were defined so that a distinction was made between documents rich in topical information (relevant and highly relevant documents) and poor in topical information (marginally relevant documents). It turned out that about 50% of documents assessed as relevant were regarded as marginal. The characteristics of the relevance corpus and lessons learned from the reassessment project are discussed. The need to develop more elaborated relevance assessment schemes is emphasized.

## Categories and Subject Descriptors

H.3.3 Information Search and Retrieval

## General Terms

Experimentation

## Keywords

Test collections, TREC, Relevance Assessments, Graded relevance, Experiment

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR'02, August 11-15, 2002, Tampere, Finland.

Copyright 2002 ACM 1-58113-561-0/02/0008...\$5.00.

## 1. INTRODUCTION

Relevance judgments are of critical importance to an information retrieval test collection. In TREC, for example, the pooling method was developed to guarantee that the list of documents assessed for relevance is as comprehensive as possible [13, 14]. The effects of variations in relevance assessments on the measured retrieval effectiveness have also been studied [10]. On the other hand, the notion of relevance has been operationalized in a very simple way. A binary scale has been used to classify documents either topically relevant or topically irrelevant. Relevance criteria have been quite liberal. The guidelines of TREC state:

*“Only binary judgments (“relevant” or “not relevant”) are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document).”* [9]

The use of binary assessments has been criticized for the lack of realism. For example, studies on Web searching have raised the need for improving the capability of IR systems to find highly relevant documents at the expense of marginally relevant ones [3]. Järvelin and Kekäläinen [5] and Voorhees [11] suggested that IR systems effective in finding highly relevant documents might suffer of binary and liberal relevance criteria. In the most recent TREC Web tracks, a three-point relevance scale has been used: *irrelevant*, *relevant* and *highly relevant* [4]. Other TREC collections have held to binary assessments.

The notion *degree of relevance* brings into play the potential usefulness of documents for a reader trying to learn about a topic. This includes an assumption that the reader has already at least an intuitive idea about the topic. This assumption is fair since we expect that the person is able to formulate a query about the topic.

For example, if the topic is *What are the applications of robotics in the world today* (TREC topic 392) we may assume that the user has a general intuition what robots are (some kind of automatic machinery) and that robots may be used for different types of applications. A document that tells that some Company X develops robot-based applications, but does not specify them, is topically relevant but does not help the user to learn about the topic. The document gave only a pointer to an external source of information (Company X).

The above example emphasizes the need to define more explicitly the criteria used in topical relevance assessments especially when the degree of relevance is considered. Earlier research has not defined clearly the criteria for multigraded relevance. The borderline between irrelevant, marginally relevant and useful documents is especially interesting.

This paper reports a project (at the University of Tampere) in which a set of 38 topics from TREC-7 and TREC-8 was exposed to graded relevance assessments. The main goal of the project was to create a subcollection where the capability of IR systems to focus on highly relevant documents could be studied. We also wanted to analyze and characterize the pools of relevant documents. The third goal was to learn more about the process and the outcome of graded relevance assessments.

Section 2 gives a general description of the reassessment project. Section 3 presents the basic characteristics of the graded relevance corpus, and its major differences compared to TREC. Section 4 sums up the main findings and presents conclusions.

## 2. DATA AND METHODS

We accepted the basic topicality assumption on which relevance assessments in the TREC collection are based. We just expanded relevance criteria to cover the degree of relevance. We attempted to follow the procedures of TREC as strictly as possible. However, some obvious differences between processes should be kept in mind when comparing results:

- TREC assessors create the topics, define relevance criteria, and compose topic descriptions by themselves. Our assessors used written topic descriptions as they had been documented by the TREC assessors.
- Our document pools contained the same relevant documents but only about 5% of irrelevant documents processed by TREC assessors.
- TREC assessors read documents on screen and could enter any keywords they liked to be highlighted in texts. We used documents printed on paper without highlighting.

We could use the original TREC assessments as a reference for the present work (but of course this information was not available for our assessors).

### 2.1 The subset of topics and documents

The topics and, especially, the way of representing the topics have varied during the history of TREC. We were encouraged to use topics from TREC-6 to TREC-8 because various reasons justified the avoidance of topics from earlier TRECs [Personal communications, Donna Harman, April 10, 2001]. For technical reasons we selected the topics 351-450 of TREC-7 and TREC-8 for further consideration.

We could not afford assessing the documents from the pools of one hundred topics, and therefore selected a subset of topics applying the following criteria:

- The number of documents assessed relevant by TREC should be more than 30 per topic. The lower limit was an attempt to keep the number of highly relevant documents adequate for testing. If the number of highly relevant documents is small, performance measures tend to become unstable making comparisons unreliable [11].

- The upper limit for the number of documents per topic was set to keep the cost of assessment reasonable. The sum of relevant documents and a 5% sample of irrelevant documents from the TREC pool should not exceed 200. (For the first 8 topics assessed, the criterion for the upper bound was slightly different, and the sample size of irrelevant documents was 10%.)
- All topics that were expected to require good background knowledge of American or British life and culture (e.g. administration or justice) were excluded. Similarly, any topic with which all assessors felt uncertain was excluded.

Some 40 topics were selected for the project but the assessment process could be completed for 38 topics, only. The pool contained 5737 documents of which 2772 had been assessed relevant in TREC. The 5% sample of irrelevant documents was selected by using the end digits of document identifiers as a sampling criterion.

### 2.2 Relevance criteria

Assessors were guided to base their work on simple assumptions about the user. The relevance should be judged considering a user needing general or specific information about a topic (for example, an author writing a review on the topic). Other aspects than topicality and the degree of topical relevance (the extent to which the text discusses the topic) should be overlooked.

The relevance of documents was assessed on a four-point scale developed originally for a Finnish test collection at the University of Tampere [7]:

- (0) The document does not contain any information about the topic.
- (1) The document only points to the topic. It does not contain more or other information than the topic description. Typical extent: one sentence or fact.
- (2) The document contains more information than the topic description but the presentation is not exhaustive. In case of a multi-faceted topic, only some of the sub-themes or viewpoints are covered. Typical extent: one text paragraph, 2-3 sentences or facts.
- (3) The document discusses the themes of the topic exhaustively. In case of a multi-faceted topic, all or most sub-themes or viewpoints are covered. Typical extent: several text paragraphs, at least 4 sentences or facts.

The basic threshold for accepting a document relevant was the same as in TREC. The criteria for *marginally relevant* items (level 1) embody so low degree of relevance that the document hardly contributes the user although it contains a piece of text mentioning the topic. Documents graded *relevant* (level 2) are expected to gain some new information to the user already having a general intuition about the topic. *Highly relevant* (level 3) documents are expected to help the user to take a good command of the topic.

Most experimenters [e.g. 1, 6, 2] as well as the TREC Web track [4] applying graded relevance scales have not made a distinction between relevant but useless (level 1) and relevant and restrictedly useful (level 2) documents. If we take a stringent standpoint, the IR system should reject marginally relevant

documents although they may contain the same words and phrases as more relevant documents.

### 2.3 Recruitment of assessors

Native speakers of English were not available for the project but, fortunately, fluent readers of English texts could be recruited. The group of six Master’s students of information studies - two of them already having a Master’s degree in another subject – was selected. All assessors had learned the basics of English at school (typically 2-4 hours per week for a period of ten or eleven years), passed the English courses at the University, and trained their reading proficiency in literature exams. In addition, individual assessors had remarkable but different experiences in practicing English: formal university studies in English or English literature, studies abroad, a long-lasting hobby of reading fiction in English, etc.

Our aim was to guarantee that the proficiency of assessors in English would not become an issue of concern in the project. As it turned out later, the specific fields of expertise needed in interpreting the themes of documents – not the language - were a problem in some topics. However, this is not a problem specific to our project. All assessors working with documents of varying themes feel uncomfortable with some documents of unfamiliar subjects.

We did not test the proficiency of assessors in English or their general or special traits formally. We knew their backgrounds and success in Master’s studies. We focused on organizing and controlling the assessment work itself. We also used TREC relevance data as a reference to identify potential problems.

### 2.4 Assessment procedure

The assessors were given written guidelines and a short introduction to the goals and tasks of the project. Next, all assessors judged a small set of documents from two topics (from TREC-6). Resulting judgments were compared and the problems observed were discussed in a meeting to clarify the details of assessment practices. Also later, regular meetings were organized for the group to discuss the problems at hand and to refine working practices through the project lasting a period of six months.

The documents shorter than one hundred pages were printed on paper for assessments. Longer documents were judged on screen. The relevance values were put down on the top of printed documents and the passages of documents that contained relevant information were also marked up.

The assessors were asked to make written notes of all interpretations they made on the description of the topic or on relevance criteria applied. They were also asked to document any problems or contradictions faced in topic descriptions. We emphasized that the assessor should clarify and fix their relevance criteria by first scanning some 20-30 documents to get a general touch on the topic and on the content of documents in hand. The assessors were advised to mark problematic documents by a question mark for a collective consideration later, if necessary.

The practices of assessment were created by three assessors during the first phase of the project. At this phase, two assessors worked on each topic to

produce parallel assessments. We could use parallel data to control the quality of work and to develop working practices. Soon we realized that the project was advancing too slowly and we could not afford duplicate assessments. Three new assessors were recruited and trained. During the second phase, each topic was treated by one assessor, only. As a result, we have two parallel assessments for eight topics and single assessments for 30 topics.

### 2.5 Follow-up

Our goal was not only to produce a corpus of graded relevance for the selected topics of TREC but also to learn about the process itself. Our basic principle was that the group of assessors is the major intellectual resource in this project, and that we focus on collecting data on their experiences. To avoid any disturbance affecting relevance assessments we did not observe the actual behavior of subjects and did not ask them to document their inference in individual assessments. The only extra task for non-problematic documents was to roughly mark up the passages of text that contained relevant information.

Later, the new relevance corpus was compared to the original relevance assessments of TREC to identify all documents where assessments diverged. The assessors were given a sample of documents where they had rated a document irrelevant although it had been rated relevant in TREC, and vice versa relevant or highly relevant although irrelevant in TREC. They were asked to analyze each document once more and reconsider if they (a) remain firm in the original assessment, (b) accept TREC assessment as the right one, or (c) change the original assessment but do not accept that of TREC. In all cases, the assessor was asked for reasons on which the decision was based.

The analysis of documents where assessments diverged was part of a larger effort to collect data about assessor experiences. For each topic an assessor had worked on, s/he was given a bundle of documents including a copy of the topic description page on which they had made their own notes, a couple of sample documents assessed relevant (to recollect the assessment process), two pairs of highly and marginally relevant documents (to explain topic related arguments how marginally and highly relevant documents differ), and the set of documents where assessments had diverged from TREC (as described above). The bundle of documents contained also a questionnaire focusing on the characteristics of the topic and the set of documents assessed.

The assessors analyzed the bundle of documents and answered the questionnaires at their own pace. After the material had been returned and analyzed, the assessors were invited to a personal interview to clarify unclear or incomplete answers in the previous data collection phase, and to discuss their experiences. The results from the post-interview are not discussed here.

## 3. RESULTS

**Table 1. Distribution of documents across relevance levels earlier assessed relevant or irrelevant in TREC and reassessed in the UTA project (38 topics, # = number of documents).**

Levels of relevance	TREC rel		TREC irrel		Total			UTA rel	
	#	%	#	%	#	%	#/topic	#	%
Rel=3	353	13 %	11	0 %	364	6 %	10	364	16 %
Rel=2	724	26 %	40	1 %	764	13 %	20	764	34 %
Rel=1	1004	36 %	134	5 %	1138	20 %	30	1138	50 %
Rel=0	691	25 %	2780	94 %	3471	61 %	91		
Total	2772	100 %	2965	100 %	5737	100 %	151	2266	100 %

The basic data on assessed documents are presented in Table 1. The total number of documents assessed was 5737 of which 2772 (48%) had been assessed relevant and 2965 (52%) irrelevant in TREC. In the reassessment, 61% of documents were found irrelevant, 20% marginally relevant, 13% relevant, and only 6% highly relevant.

On the average, 60 more or less relevant documents are available per topic of which 10 were highly relevant, 20 relevant and 30 marginally relevant. About 50% of documents assessed relevant were considered marginal, one third relevant, and only 16% highly relevant (Table 1, the last column).

Table 1 (the first column) shows that 691 documents rated relevant in TREC were rated irrelevant in our reassessments (25% of relevant in TREC). We call this deviation as *type A inconsistency*. In addition, 1004 documents relevant in TREC (36%) were considered as marginally relevant by our assessors. The figures suggest that the degree of relevance seems to be quite low in over 60% of documents assessed relevant in TREC. However, before making conclusions we need to analyze further inconsistently assessed documents and the arguments on which our assessors based their decisions. Some relevant documents were either missed or assessed differently.

Another result was that our assessors found some relevant documents that had not been identified, or considered as relevant by the TREC assessors. We call this deviation as *type B inconsistency*. The number of relevant and highly relevant documents missed or judged irrelevant by TREC was very small (about 1% of all assessed irrelevant in TREC). However, our evidence on the missed relevant documents is weak since our sampling method for irrelevant documents was not directed to hunt the missed ones. This problem has been discussed and

**Table 2. The effect of asking assessors to reassess type A inconsistently judged documents (a sample of 69 out of 691 documents).**

Relevance level	1st time		2nd time	
	#	#	#	%
Rel=3			0	0 %
Rel=2			4	6 %
Rel=1			21	30 %
Rel=0	69		44	64 %
Total	69		69	100 %

**Table 3. The effect of asking assessors to reassess type B inconsistently judged documents (all relevant or highly relevant documents).**

Relevance level	1st time		2nd time	
	#	#	#	%
Rel=3	6		5	11 %
Rel=2	41		30	62 %
Rel=1			2	4 %
Rel=0			10	22 %
Total	47		47	100 %

experimented by Zobel [14].

### 3.1 Type A inconsistency

A sample of 69 documents (about 10%) regarded as relevant in TREC but rated as irrelevant in our assessments (type A inconsistency) was sent to the assessors for post-analysis. The sample was created using the following criteria: 1) If at least one inconsistently assessed document was available for a topic, it should be included. 2) No more than three documents were selected per topic. 3) All documents should be from different sub-collections if possible. 4) All documents longer than 5 pages were excluded.

The aim of criteria 1-3 was to guarantee that the sample was representative in terms of topics, assessors, and document types. The fourth criterion was used not to overload and demotivate our assessors at a critical stage of data collection.

The results of the second assessment are presented in Table 2. The assessors did not change their ratings for two thirds of documents (64%), accepted 30% as marginally relevant, and 6% as relevant. None of the documents reassessed were judged highly relevant. This finding suggests that we can trust that our assessors were quite reliable in identifying highly relevant documents at least if we believe on their criteria and arguments in the second assessment.

Based on 4 unrecognized relevant (rel=2) documents from the sample we may estimate that about 40 relevant (rel=2) documents out of the 691 documents judged relevant by TREC assessors may have been missed by our assessors. This is about 0.6% of all documents assessed and would add, on average, one extra relevant document per topic to the relevance corpus. Adding estimated misses would increase the share of rel=2 documents slightly from 26% to 28% (the third column in Table 1).

The share of missed marginally relevant documents is quite high in the sample (30%). We may estimate that about 210 marginally relevant documents were regarded as irrelevant by our assessors. This finding suggests that the share of marginally relevant documents is even higher (about 44% instead of 36%) than expressed in Table 1. On the other hand, the number of irrelevant documents would decrease from 691 to some 440 documents, and the share from 25% to 16%, respectively.

Above figures suggest that our assessors were less effective in identifying marginally relevant documents than their TREC colleagues. This difference is probably associated with differences in the assessment practices. Our assessors read printed documents while their TREC colleagues scanned documents on screen by a browser highlighting topic related keywords. Documents where the topic is mentioned in one sentence only are easily missed in traditional scanning of text on paper.

To sum up, the set of 691 documents judged irrelevant by our assessors (type A inconsistency) is estimated to consist of 1) missed relevant and 2) differently judged documents:

- Highly relevant documents were recognized reliably.
- About 6% of relevant, and 30% of marginally relevant were missed.
- About 64% of documents were judged differently.

We will analyze the last group and reasons for different judgments in Section 3.3.

### 3.2 Type B inconsistency

A set of 185 documents that had been assessed irrelevant in TREC but relevant (at levels 1-3) by our assessors (type B inconsistency, Column 4 in Table 1). Of these, 47 were rated relevant or highly relevant and we asked our assessors to reassess them. Marginally relevant documents were excluded since we considered their role less important. Table 3 presents the results of the second assessment. It turned out that our assessors accepted TREC relevance ratings at least partially (rel=0 or 1) for 12 documents (26%) but defended their original judgment for 35 documents (74%).

The rate of missing relevant or highly relevant documents by TREC was about 0.6% of all documents in the sample. Practically, there was no difference in the rate of missing relevant and highly relevant documents between TREC and our assessors.

### 3.3 Why documents were assessed differently?

Documents where assessments differed between TREC and us were not evenly distributed over the 38 topics. 90% of documents which our assessors had rated irrelevant contrary to TREC came from 16 topics. 90% of relevant and highly relevant documents rated irrelevant by the TREC assessors were related to 9 topics. This is not a surprise since the clarity of relevance criteria varied from topic to topic as the analysis of assessor experiences revealed.

The assessors were asked to judge the clarity of descriptions for each topic they had evaluated (4-point scale: 1. very ambiguous ... 4. clear), and their general confidence on decisions concerning each topic (scale: 1. lower than typical, 2. typical, 3. higher than typical). We were interested to see if assessor perceived clarity and confidence was associated with the variation of consistency with TREC assessments.

The topics were sorted in order of descending number of documents assessed irrelevant although regarded as relevant in TREC (type A inconsistency). The sorted list of topics was divided into quartiles containing 9, 10, 10, and 9 topics. The averages for perceived topic clarity and confidence of decisions were calculated for each quartile of topics. The upper quartile of topics yielded 67%, the second 29%, and the third 3% of type A inconsistent documents.

**Table 4. Topic ambiguity and confidence of decisions in topic groups organized in the descending order of inconsistency instances (Type A: TREC relevant -> irrelevant; Type B: TREC irrelevant -> relevant or highly relevant).**

Topics	n	Clarity 1..4		Confidence 1..3	
		Type A	Type B	Type A	Type B
I Quartile	9	2,3	2,3	1,9	1,9
II Quartile	10	2,9	2,6	2,0	2,3
III Quartile	10	2,8	2,7	2,2	1,9
IV Quartile	9	2,8	3,2	2,3	2,4

Similar grouping of topics was made on basis of documents assessed relevant or highly relevant although irrelevant in TREC (type B inconsistency). However, the low number of type B documents (only 39) limits the reliability of this comparison.

For type A inconsistency, the average perceived clarity of topics was clearly lower for the first quartile of topics (2.3) than for topics in other quartiles (2.8-2.9). Similarly, the average confidence of decisions tended to be low for this quartile of topics (1.9 vs. 2.0-2.3). In type B inconsistency, trends were similar but the figures were not so steady because of the smaller number of topics where type B inconsistency occurred.

The figures suggest that the consistency of assessments is difficult to achieve if the assessors feel topic descriptions ambiguous. This is also associated with the confidence of decisions. In the personal interviews, assessors complained that for many topics the description (including the title) and the narrative presented contradicting relevance criteria or otherwise did not give a solid basis for relevance assessments. Our assessors had to specify relevance criteria by themselves to guarantee consistent treatment of documents. However, new criteria were probably not always in line with the intentions of original assessors.

### 3.4 Economics of graded assessments

Judging document relevant is a quite straightforward process if liberal criteria are applied. The assessor only needs to identify one sentence or a text paragraph where the topic is discussed. In graded assessments, some extra work is required to specify the degree of relevance for documents found relevant. If the relevance of a document cannot be identified easily, the treatment of documents is the same in binary and graded assessments. The assessor is expected to verify the irrelevance by reading the whole document.

In all, the total time needed to assess slightly more than 7000 documents (parallel assessments included) required approximately 78 full workweeks (about 36 hours each). This figure includes training, meetings; printing of documents, data entry, vacations, etc. but excludes the work done by supervisors.

The average number of 73 documents was processed per week and per person during the first phase of the project. The speed rose up to 105 documents per week in the second phase. During the second phase, the practices were already fixed and more time could be allocated for actual assessments.

Individual differences in effectiveness were notable and having enough practice in assessments seemed to increase output. During the second phase, one of the experienced assessors reached an average rate of 220 assessed documents per week (about 6 documents per hour) but none of the newcomers could exceed 130 documents per week. The figures suggest that pre-testing of assessors is worth considering in major assessment projects.

## 4. CONCLUSIONS AND DISCUSSION

### 4.1 Binary and graded assessments

One major contribution of this study is based on the distinction between documents only pointing to the topic (rel=1, marginal) and those really containing some potentially useful information (rel=2, relevant). About 50% of the documents assessed as relevant were marginal. The other half of documents contained potentially useful information to the user. Only 16% of the relevant documents were highly relevant.

The share of documents rated as irrelevant or marginal in the TREC relevance corpus was even higher (about 60%). The figure is biased by the differences in the assessment processes between us, and TREC. Voorhees [12] found in a parallel test that secondary assessors working on the basis of written topic descriptions tend to reject a notable share of documents that the author of a TREC topic (primary assessor) regards as relevant. On the other hand, Voorhees [12] observed that secondary assessors found very few new, relevant documents. Our data was in line with this finding (6%).

Anyhow, the results show that the relevance threshold is quite low in TREC. This is quite natural if we consider the themes of discussions in the past: the concern about missed relevant documents [14] together with liberal relevance criteria and binary scaling [9, 13]. Binary relevance has been a convenient approach in test collections since it makes performance calculations simple, keeps the cost of assessments low, and maximizes the number of relevant documents per topic guaranteeing the stability of measures.

This is all good, but we have to ask: are we able to conduct experiments that we consider as meaningful? Without a doubt, TREC is an indispensable tool for studying retrieval systems aimed at high-recall searching. However, we definitely need test collections for IR systems aimed to retrieve highly relevant documents effectively from large databases (e.g. Web indexes). This requires that we build test collections that provide more differentiated relevance data.

## 4.2 Assessment costs and procedures

Can we afford the higher cost of graded relevance assessments? Our project was not optimized in terms of costs and the volume of throughput. The productivity of assessments can be substantially increased by using appropriate browsing tools, developing the routines, and recruiting the most effective assessors. Obviously, the costs are higher for graded assessments than for binary assessments. However, one might imagine that the assessment cost when using our 4-point scale does not differ much from the cost when using the 3-point scale of the Web Track. This comparison suggests that if the costs are manageable in the Web Track, they can be afforded when using the 4-point scale.

In TREC, browsers highlighting topic related keywords in the document help one locate a text of interest. The use of this tool could be improved by developing the procedure of how and how exhaustively keywords are composed. Relevant and highly relevant documents contain a larger variety and number of topic related terms, [see 8]. It is likely that they can be identified quite reliably if the highlighting procedure is designed well.

In a test collection based on graded relevance assessments, it is a minor issue if a small share of marginally relevant documents is missed. This fact raises the question: could the pooling techniques be developed to decrease the total number of documents assessed?

The easiest way to resolve the need for test collections providing a multigraded relevance corpus is to reassess the pools of selected TREC topics. It is not likely that TREC could afford multigraded relevance assessments instead of binary ones as an annual routine. The costs of reassessment can be limited, because most (if not all) of the irrelevant documents can be ignored.

## 4.3 Performance measures

The number of highly relevant documents per topic tends to be small. As Voorhees [11] pointed out, results based on small document sets may be unstable. Thus, she recommended the use of the Discounted Cumulative Gain (DCG) method proposed by Järvelin and Kekäläinen [5] to combine and weight the evidence from different levels of relevance. The advantage of the 4-point relevance scale is that relevant and useful (rel=2) and relevant but potentially useless (rel=1) documents can be weighted differently. The problem of small sets can be reduced, if the difference of weights between highly relevant (rel=3) and relevant (rel=2) is kept modest to increase the stability of measures. On the other hand, marginally relevant documents may be given a very low, if any, weight. Another method to increase stability is to have more topics [8].

The three-point relevance scale launched in the Web Track of TREC is only a small step forward since the distinction between topically relevant but potentially useless (rel=1) and topically relevant and potentially useful (rel=2) has not been made. The three-point scale gives fewer choices in the use of weights.

## 4.4 Problems and future research

The ultimate test for the usefulness of graded relevance scales is: does it matter in practice if we use graded relevance instead of binary relevance? This could be done by using old TREC data to investigate the effect of graded relevance scales on measured performance differences. If the results are clearly different when graded relevance is used, the extra effort required and resources consumed may be justified. Voorhees [11] did a test of this kind to find out the potential benefits of the Discounted Cumulative Gain (DCG) method.

The use of old TREC data is necessary only if we want to verify a hypothesis that the findings in the past TREC experiments were biased by the liberal relevance assessments. We expect that liberal binary assessments are ok for experiments focusing on high recall searching. Our main point is that we need multigraded relevance data to study retrieval phenomena associated with the degree of relevance.

The set of 38 topics is small for retrospective analysis, and even that set is divided among TREC-7 and TREC-8. We found more reasonable to design new experiments in which research questions deal with retrieving highly relevant documents. This experiment is in progress. The relevance corpus will become available to the TREC community after the data has been checked in course of the experiment.

Apart from actual tests, more research is needed to develop graded relevance scales. For example, the benefits of using a 4-point scale instead of a 3-point scale together with a higher basic threshold, is an open question. If we do not require that all marginally relevant documents need to be found and pooled, the goals and practices of relevance assessment processes have to be reconsidered.

Our project applied an empirical approach to emphasize the methodological challenge of developing relevance assessments in test collections. This was a modest but determined step towards realism in relevance assessments as the degree of relevance was taken into account. The next natural step could be to develop relevance schemes than can be used to handle overlaps between document contents. Any static characteristic of a document is a

potential search criterion and also a potential target for relevance assessments in test collections.

## 5. ACKNOWLEDGMENTS

The author is very grateful to Professors Jaana Kekäläinen and Kal Järvelin for their effort in the project. We all thank Donna Harman for her encouraging advice. I am also very grateful to the group of referees for their constructive criticism and comments.

Otto Auranen, Jussi Koivisto, Marko Panttila, Mira Roine, Reetta Saine and Mikael Vakkari made an enormous assessment work and were actively developing the procedure of assessment.

The members of the FIRE group have also helped to improve the manuscript. Especially, the author thanks the visiting colleagues in FIRE, Diane Sonnenwald and Per Ahlgren, for their valuable comments.

## 6. REFERENCES

- [1] Blair, D.C. and Maron, M.E. An evaluation of retrieval effectiveness for a full-text document retrieval system. *Commun. ACM* 28, 3, 289-299.
- [2] Burgin, R. Variations on relevance judgments and the evaluation of retrieval performance. *Information Processing & Management* 28, 5, 619-627,
- [3] Gordon, M. and Pathak, P. Finding information on the world wide web: The retrieval effectiveness of search engines. *Information Processing and Management*, 35, 2, 141 – 180.
- [4] Hawking, D. Overview of the TREC-9 Web Track. <http://trec.nist.gov/pubs/trec9/papers/web9.pdf> [Cited 2 January 2002].
- [5] Järvelin, K. & Kekäläinen, J. IR evaluation methods for highly relevant documents. In: Belkin, N.J., Ingwersen, P. and Leong, M.-K. (eds.) *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York 2000, 41-48.
- [6] Saracevic, T., Kantor, P. et al. A study of information seeking and retrieving. I Background and methodology. *Journal of the American Society for Information Science* 39, 3, 161-176.
- [7] Sormunen, E. A method for measuring wide range performance of Boolean queries in full-text databases. Doctoral Thesis. University of Tampere. *Acta Electronica Universitatis Tamperensis* 34, URL: <http://acta.uta.fi/teos.phtml?3786>, Tampere 2000.
- [8] Sormunen, E., Kekäläinen, J., Koivisto, J. & Järvelin, K. Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. *Journal of Documentation* 57, 3, 358-374.
- [9] TREC homepage, Data – English relevance judgements. Available at: [http://trec.nist.gov/data/rejudge\\_eng.html](http://trec.nist.gov/data/rejudge_eng.html) [Cited 31 December 2001].
- [10] Voorhees, E. Variations in relevance judgments and the measurement of retrieval effectiveness. In: Croft, W.B. et al. (eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, August 1998). ACM Press, New York., 315-323.
- [11] Voorhees, E. Evaluation by highly relevant documents. In: Croft, W.B. et al. (eds.). *Proceedings of the 24<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, September 2001). ACM Press, New York, 74-82.
- [12] Voorhees, E. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management* 36, (2000), 697-716.
- [13] Voorhees, E. & Harman, D. Overview of the Seventh Text Retrieval Conference (TREC-7). 1999. [http://trec.nist.gov/pubs/trec7/papers/overview\\_7.pdf.gz](http://trec.nist.gov/pubs/trec7/papers/overview_7.pdf.gz) [Cited 2 January 2002].
- [14] Zobel, J. How reliable are the results of large-scale information retrieval experiments? In: Croft, W.B. et al. (eds.), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, August 1998). ACM Press, New York.

