Relevance Judgments between TREC and Non-TREC Assessors

Azzah Al-Maskari
lip05aaa@shef.ac.uk

Mark Sanderson
m.sanderson@shef.ac.uk

Paul Clough
p.d.clough@shef.ac.uk

Dept. of Information Studies
Sheffield, S1 4DP, UK
University of Sheffield

ABSTRACT

This paper investigates the agreement of relevance assessments between official TREC judgments and those generated from an interactive IR experiment. Results show that 63% of documents judged relevant by our users matched official TREC judgments. Several factors contributed to differences in the agreements: the number of retrieved relevant documents; the number of relevant documents judged; system effectiveness per topic and the ranking of relevant documents.

Categories and Subject Descriptors

H.3.0 [**Information Storage and Retrieval]:** *General*
General Terms: Measurement, Human factors

Keywords: user study, TREC Relevance assessment

1. INTRODUCTION

Relevance assessments are of critical importance to the evaluation of information retrieval systems. The Text REtrieval Conference[1] (TREC) has established an evaluation practice where a binary relevance scale is combined with liberal relevance criteria. However, the low threshold for relevance criteria followed in TREC has been criticized in affecting the ability to identify and develop IR methods capable at retrieving highly relevant documents [1]. Voorhees [5] studied the effect of variations in relevance assessments on the measured retrieval effectiveness. She found 32.8% agreement between official TREC judges and a set of new assessors. Despite this apparently low number, she concluded that different relevance assessments did not affect the ranking of systems in the TREC evaluation. Recent studies have examined graded relevance: Sormunen [2] used a four-point scale to compare agreement between relevance judgments and to reassess document pools for 38 TREC topics. He found 39% of documents rated relevant by TREC assessors were similarly rated relevant by a set of new assessors. Vakkari & Sormunen [4] explored the consistency with which users could identify relevant documents. They found that users could identify 45% of documents previously judged relevant by TREC assessors. Similarly, a recent study by Turpin, & Scholer [3] showed 45% agreement between TREC assessment and a group of student users. All of these previous studies [2] [4] [5] are based on re-assessing documents pools of TREC topics.

In this study, we examine the overlap in agreement between official TREC assessments and the relevance judgments created by new users within an Interactive IR experiment. The overlap is defined as the size of intersection between the sets of documents judged relevant by TREC assessors and our users, divided by the size of set of documents users judged as relevant. In this paper, we also explore circumstances in which users strongly disagree with official TREC assessments.

SIGIR' 08, July 20-24, 2008, Singapore.

1. EXPERIMENTAL DESIGN

Fifty-six participants were recruited to engage in a search task that required saving as many relevant documents as possible for a set of 56 TREC topics. An interactive search system[2] retrieving over the TREC-8 document collection was used. Sets of 8 users completed a search for the same set of 8 topics. Users were given 7 minutes for each topic, and a Latin-Square arrangement utilised to distribute the order of the topics amongst users (to reduce the effects of topic order on results). Users were presented with the description and narrative fields of TREC topics as information needs to be satisfied. They were free to issue multiple queries for each topic within the 7 minutes. The narrative field served as guidance on assessing document relevance using a ternary relevance scheme: highly relevant[3], partially relevant[4] or not relevant.

2. RESULTS

Of the 2,262 documents judged relevant by our users (R(U)), 1,428 (63%) were also assessed relevant in TREC (we also label it as *consistent judgment*); while 834 (37%) were assessed irrelevant (NR(T)) in TREC (we also label it as *inconsistent judgment*) as shown in Table 1. Of the 1,428 consistent judgments, users rated 37% of the documents as partially relevant and 63% as highly relevant; while of the 834 documents they rated 61% as partially relevant and 39% as highly relevant. Out of the 2,133 irrelevant judgments (NR(U)) made by our users, 308 (14%) documents were regarded as relevant in TREC.

Table 1: The relevance categories of retrieved documents.

		TREC assessors		
Users		R (T)	NR (T)	total
	R (U)	1428	834	2262
	NR (U)	308	1825	2133

The agreement between our users and TREC assessors was measured with respect to document rank (Table 2). We found a higher level of agreement between relevance judgments between our users and TREC assessors for documents ranked highly, and a lower level of agreement for documents ranked lower. The highest

[1] http://trec.nist.gov/ [site accessed: 22/02/08]

[2] http://www.info.uta.fi/julkaisut/pdf/qparn1.pdf [site accessed: 22/02/08]

[3] The document directly addresses the core issue of the topic.

[4] The document only points to the topic: it does a not discus the themes of the topic thoroughly.

agreement was at Rank 1 and it decreased as the rank increased. Relevance judgments where assessments differed between TREC assessors and our users were not evenly distributed over the 56 topics and the 56 users.

Table 2: Agreement at various rank positions.

Rank	1	5	10	20	50	100
Consistent judgment	74%	71%	67%	65%	64%	63%
Irrelevant judgment[5]	13%	13%	14%	14%	14%	14%
Overall agreement[6]	82%	79%	76%	75%	75%	75%

Eighty percent of documents that our users had rated irrelevant contrary to TREC came from 25 topics and 24 users, while 80% of documents that our users had rated relevant contrary to TREC came from 29 topics and 33 users. We have examined whether the same users who judged documents relevant (contrary to TREC assessment) were the same who judged documents irrelevant (contrary to TREC assessment). This was to investigate whether the same users constantly disagreed with TREC judgments, however, the resulting correlation did not support this assumption ($r=-0.16$, $p<.000$). Similarly, we investigated the same effect based on individual topics (whether users judged documents relevant and irrelevant contrary to TREC assessment occurred in the same set of topics). According to the resulting correlation ($r=-0.13$, $p<0.000$) topics did not have an effect on users' disagreeing with TREC assessment.

System effectiveness (as measured by MAP), per topic, did effect users' agreement with TREC assessors. It was found that users consistent judgements with TREC assessors correlated ($r=0.4$) significantly more ($p=0.00$) with system effectiveness than their inconsistent judgements ($r=-0.06$). The low correlation between inconsistent judgements and system effectiveness indicated that users' disagreement with TREC judgment increased as the system effectiveness decreased. For example topic 302 with MAP=0.55, total documents judged relevant by our users were 11, 10 of them were judged consistent with TREC assessment (relevant) whereas only one document was judged inconsistent with TREC. On the contrary, topic 420 with MAP=0.11, total documents judged relevant by the users were 12, 9 of them were regarded irrelevant by TREC assessors and only three documents were judged consistently relevant with TREC assessment. This meant that people disagreed more with TREC judgements on topics with lower system effectiveness, and consequently agreed more on the topics with higher system effectiveness.

We further analyzed possible causes for disagreements between our users and the TREC assessors. We found that as the total number of retrieved relevant documents for a particular topic decreased users tended to select proportionally more documents inconsistent with TREC judgments ($r=0.02$, $p<.000$), whereas users tended to rate documents more consistently with TREC assessments ($r=0.65$) as the total number of retrieved relevant documents increased. For example, topic 410, the total number of retrieved relevant documents was 9, users judged 4 documents inconsistently relevant with TREC while only 1 document was judged consistent with TREC and the rest were not identified by the users. On the other hand, topic 415, the total number of retrieved relevant documents was 34, users judged 13 documents consistently relevant with TREC while only 3 documents were judged inconsistent with TREC. Thus, a small number of potentially relevant documents in the retrieved sets lead users to accept more documents inconsistently relevant with TREC assessment. We also found that as the number of relevant documents judged increased (consistent and inconsistent with TREC), users tended to rate documents more consistently with TREC assessments ($r=0.83$) as compared to the inconsistent ratings ($r=0.63$). For example topic 415 mentioned above, users judged in total 16 documents, the majority of them were consistent with TREC (13). This suggested that if users judged few relevant documents, they may have felt obliged to continue saving further documents (relevant or irrelevant with regards to TREC assessments) due to the fear of "failure" to complete the set task. For example, topic 356, the total number of relevant documents judged by the users was 5; only 1 document was consistently relevant with TREC judgments while 4 were inconsistent.

The lack of inconsistency in relevance assessments between our users and TREC assessors stemmed from the following conditions: users who retrieved few documents and judged few relevant documents appeared to relax their criteria for relevance, accepting more documents as relevant contradicting TREC assessments. Moreover, some users had difficulty identifying relevant documents because they simply found some TREC topics harder than others. This signifies that we are likely to get lower agreement in relevance judgments for topics that relevant documents are difficult to find. Therefore, in interactive IR studies which make use of TREC test collections, when these conditions occur, care should be taken when comparing user effectiveness with system effectiveness.

4. CONCLUSION

Results show that 63% of documents judged relevant by our users matched official TREC judgments. One explanation for this agreement could be that TREC topics used in this experiment (and associated relevance) were clear and lack ambiguity. This high agreement might indicate that when a retrieval system believe documents are relevant, human are also likely to agree on relevance. These findings help us understand the potential impact of using interactive studies to generate a test collection.

5. ACKNOWLEDGMENTS

We would like to thank Ministry of Manpower, Oman, and the TrebleCLEF project (IST-FP7- 215231) for funding this study.

6. REFERENCES

[1] K. Järvelin & J. Kekäläinen. IR evaluation methods for retrieving highly relevant documents. *SIGIR*. p.41-48, 2000

[2] E. Sormunen. Liberal relevance criteria of TREC: Counting on neglible documents? *SIGIR*. Pp.324–30.2002

[3] A. Turpin & F. Scholer. User Performance versus Precision Measures for Simple Search Tasks. *SIGIR*. PP.11-18. Washington, USA. 2006

[4] P. Vakkari & E. Sormunen. The influence of relevance levels on the effectiveness of interactive information retrieval. *JASIS*. 55(11), 963–969.2004

[5] E.M.Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management. 36(5), 697–716. 2000.

[5] Users rated documents irrelevant while they are relevant in TREC.

[6] Docs judged relevant & irrelevant by both TREC assessors & users.