

Quantative Analysis of The Impact of Judging Inconsistency on The Performance of Relevance Feedback

Xiangyu Jin
University of Virginia
151 Engineer Way
Charlottesville VA 22903, USA
xj3a@cs.virginia.edu

James French
University of Virginia
151 Engineer Way
Charlottesville VA 22903, USA
French@cs.virginia.edu

Jonathan Michel
Science Applications
International Corporation
Charlottesville VA 22911, USA
Jonathan.D.Michel@saic.com

ABSTRACT

Practical constrains of user interfaces makes the user's judgment (during the feedback loop) deviate from her real thoughts (when full document is read). This is often overlooked in evaluation of relevance feedback. In this poster, we quantitatively analyze the impact of judging inconsistency on the performance of relevance feedback.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Relevance Feedback

General Terms

Performance, Experimentation, Measurement

Keywords

Relevance Feedback, Judging Inconsistency, Performance Evaluation

1. INTRODUCTION

Relevance feedback has been historically proven to be effective for information retrieval [5] [7]. Since it is very costly to do user-in-the-loop evaluation, large scale evaluation of relevance feedback usually employs machine simulated users instead of real human users. Under such experiments, the surrogates generated by the retrieval system (e.g., judging whether a provided document is relevant by its title)¹ are answered by machine based on the pre-defined groundtruth. This infers "perfect consistency" between relevance judgment during feedback loop and relevance judgment after the full document is read when deciding the groundtruth.

However, it is very hard to achieve such perfect consistency in practice. On one hand, practical constrains, such

¹In the following we focus our study on document-level feedback.

as time spent for judging, the screen resolution, etc., could restrict the information delivery to the user during the feedback loop. For example, in text retrieval, we can only provide a document's title, key terms, or abstraction to the user instead of providing the full document. On the other hand, the real human user, can be inconsistent with themselves. When the user learns more about her information need during the feedback process, she might change her criteria of relevance in subtle ways. What she considers relevant during feedback loop may not be considered relevant after the retrieval process terminate. Neglecting such judging inconsistency will result in exaggerated performance gain for relevance feedback. To date the impact of such inconsistent judgments on the performance of refined search has not been carefully studied. In this paper we explore the answers to the following research questions.

How often does such judging inconsistency happen? The answer to this question is related to specific user interface and retrieval environment. Unfortunately, it is unrealistic to enumerate all possible user interfaces and perform large scale tests over various environments since human subjects are involved. We focus on case studies of recent years TREC's HARD tracks in this paper, where reasonable user interfaces (the clarification forms) are designed and reasonable relevance judgments are made (by NIST assessors). Such case studies can help us better understand the state of the art.

How does such judging inconsistency affect the refined retrieval performance? If such judging inconsistency is inevitable, we want to know how it will affect the refined search performance. To study this question can help us estimate whether a relevance feedback technique would indeed help a retrieval application in practice. Moreover, relevance feedback algorithms perform similar with "perfect user" may perform quite different with real user. In this sense, some relevance feedback algorithms are more robust.

In this paper, we make an initial attempt to study the above two questions.

2. CASE STUDIES

TREC (Text Retrieval Conference)'s HARD (High Accuracy Retrieval from Documents) track² provides us an opportunity to quantitatively analyze such judging inconsistency and their impact on performance of relevance feedback. HARD provides large-scale centrally-administered evaluations for retrieval systems which allow one round of in-

²<http://trec.nist.gov/tracks.html>

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '06, August 6–10, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-59593-369-7/06/0008 ...\$5.00.

teraction. Basically HARD split the retrieval process into three phases: baseline, clarifying, and final. Initially, each participant generate their baseline runs by performing traditional ad-hoc retrievals over the HARD corpus. In the second phase, “clarification forms” (CF) are generated for each topic. These CFs are submitted for NIST assessors. Later, the filled in CFs are returned to the HARD participants to generate their refined search results (final runs). Although HARD is not designated for relevance feedback task, it is an ideal environment to study the previous two questions. First, HARD is based on large scale evaluation (about 1M documents) and upon metrics historically proven to be effective in past TRECs. Second, the assessor who answers the CFs for a specific topic is the one who decides its groundtruth. This eliminates inconsistency among different human subjects. Third, practical constrains are imposed on CFs. For example, each CF must be filled in within 3 minutes. Fourth, the assessor answers the CFs independently of the participants. This is extremely important in keeping the bias of the interface developers from creeping into the evaluation. Finally, the CFs and their judging results are available for research purpose.

In the following, we choose UIUC’s HARD 2003 submission (2 CFs) [6] and SAIC’s HARD 2005 submission (1 CF) [2] for analysis because these are the CFs we currently know the association between the surrogates (on CFs) and the documents they on behalf of. The settings are listed in Table 1.

Table 1: CF Settings

Name	ILUC-1,2	SAIC1
# Surrogates	6	8
Time Limit	3 minutes	3 minutes
Display Content	Abstraction	Keywords, Title, and Abstraction
User Interface	Showing directly	Showing when mouse over title
Choices	Relevant/Irrelevant	Relevant/Perhaps /Irrelevant
Default	Irrelevant	Perhaps

We find that the relevance judgment during feedback loop is not fully consistent with relevance judgment after the full document is read for the same user. 19.6% documents of ILUC-1, 21.7% of ILUC-2, and 22.9% of SAIC1 are judged inconsistently (if exclude those leave as “perhaps”). The detailed results are listed in Table 2.

Table 2: Judging Inconsistency (the * indicate the inconsistent part)

CFs	Groundtruth	Judged-Rel	Judged-Irrel	Perhaps
ILUC-1	Rel	28.7%	*6.3%	—
ILUC-1	Irrel	*13.3%	51.7%	—
ILUC-2	Rel	36.3%	*4.7%	—
ILUC-2	Irrel	*17.0%	42.0%	—
SAIC1	Rel	25.3%	*9.8%	12.3%
SAIC1	Irrel	*7.8%	34.0%	11.0%

Afterward, we analyze how the judging inconsistency will affect the performance of relevance feedback. We compare the refined search performance when the CFs are judged

by different users in HARD 2005 environments, including a perfect user (judging by groundtruth), a blind user (judging everything as relevant), a real user (judging results of SAIC1), and a simulated user (who randomly make 30% of its judgments inconsistent with the groundtruth) We use a BM25-ranked retrieval system to generate the baseline search result. Top 20 terms from each judged relevant document are extracted and combined to the initial query by Rocchio [4] method. We only consider positive relevance feedback at this time. Mean average precision (MAP) is reported as the evaluation metric. In order to fairly compare the results, rank shifting [3] is employed both for the baseline and refined search results (for the documents listed for judging, move relevant ones to the head of the result list and irrelevant to the end of the result list). The results are shown in Table 3.

Table 3: Performance of Relevance Feedback

Run Name	MAP	Inconsistent Rate
Baseline	0.2580	—
Perfect User	0.3629	0
Blind User	0.2953	0.5275
Real User	0.3069	0.2975
Simulated User	0.3064	0.2937

Interestingly, we find that relevance feedback in practice performs much lower than it should be. If the user can judge the document consistently with the groundtruth, the refined MAP will reach 0.3629, which is much higher than the baseline 0.2580. However, the performance of relevance feedback with a real user is around 0.30, which is comparable to the performance pseudo-relevance feedback. This indicates the bottle neck for current relevance feedback applications resides in judging inconsistency but not the relevance feedback algorithm. With better user interfaces and more appropriate information delivered, relevance feedback can be greatly improved. Moreover, our simulated user is a more appropriate estimation of relevance feedback’s performance in practice than the perfect user.

3. REFERENCES

- [1] C. W. Cleverdon. The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–192, 1967.
- [2] X. Jin, J. French, and J. Michel. Saic and university of virginia at trec 2005: Hard track. In *TREC*, 2005.
- [3] X. Jin, J. C. French, and J. Michel. Toward consistent evaluation of relevance feedback approaches in multimedia retrieval. In *Adaptive Multimedia Retrieval*, pages 191–206, 2005.
- [4] J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, 1971.
- [5] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41(4):288–297, 1990.
- [6] X. Shen and C. Zhai. Active feedback - UIUC TREC-2003 HARD experiments. In *TREC*, pages 662–666, 2003.
- [7] H. Zhang and Z. Su. Relevance feedback in CBIR. In *VDB*, pages 21–35, 2002.