User Satisfaction Task: A Proposal for NTCIR-7

Tetsuya Sakai
NewsWatch, Inc.
sakai@newswatch.co.jp

Abstract

Good test collections, coupled with good evaluation metrics, are very useful for evaluating Information Access systems efficiently. But useful to whom? The in vitro *(or* Cranfield*) evaluation paradigm has been criticised, mainly because of the absence of the user. On the other hand, user-in-the-loop evaluations are expensive, unrepeatable and often inconclusive. In light of this, we propose a new task for NTCIR that aims to directly measure the correlation between user satisfaction and evaluation metric values. To this end, we plan to reuse NTCIR-5 and NTCIR-6 Japanese monolingual newspaper test collections from the crosslingual task. Our final goal is to design new evaluation metrics that accurately approximate user satisfaction scores.*

Keywords: *user satisfaction, evaluation metrics, test collections.*

1 Motivation and Background

Good test collections, coupled with good evaluation metrics, are very useful for evaluating Information Access systems efficiently. But useful to whom? The *in vitro* (or *Cranfield*) evaluation paradigm has been criticised, mainly because of the absence of the user. For example, recently at ACM SIGIR 2006, Turpin an Scholer [14] reported that Mean Average Precision (MAP), *the* most widely-used IR evaluation metric, is not significantly correlated with the time the user requires to identify one relevant document[1]. At a workshop immediately following that conference, "Death to Average Precision" was discussed [4]. On the other hand, user-in-the-loop evaluations are expensive, unrepeatable and often inconclusive. Voorhees [15], for example, defends the test collection paradigm while pointing out these weaknesses of existing user-based studies.

Within the test collection paradigm, many researchers have proposed new evaluation metrics [2, 3, 5, 8, 9, 17] and/or have addressed the problem of *evaluating evaluation* using test collections and IR metrics [7, 11, 12, 16]. Are any of these studies relevant at all to the real user environment? Many researchers *believe* at least some of them are relevant and useful, but they have no substantial proof. In particular, researchers optimise their IR systems using a few evaluation metrics of their choice in the hope of improving user satisfaction, but they do not really know whether relying on these metrics is a very good idea, or, even if they are, how much improvement is practically visible to the user.

At the aforementioned SIGIR 2006 Workshop [4], we proposed that *in vitro* evaluations should be done (say) 80% of the time, but user-in-the-loop evaluations should be done (say) 20% of the time so that the consistency of the two evaluation paradigms can be verified periodically. In light of this, we propose a new task for NTCIR that aims to directly measure the correlation between user satisfaction and evaluation metric values. To this end, we plan to reuse NTCIR-5 and NTCIR-6 Japanese monolingual newspaper test collections from the crosslingual task. Our final goal is to design new evaluation metrics that accurately approximate user satisfaction scores.

2 Tentative Task Definition

On the surface, the user satisfaction task ("USAT") is just like any other *ad hoc* monolingual document retrieval task. We plan to reuse the NTCIR-5 and NTCIR-6 Japanese monolingual newspaper test collections from the crosslingual (CLIR) task, so the "right answers" are known in advance to all participants. Figure 1 shows how the NTCIR-3, 4, 5 and 6 Japanese test collections are related to one another: As the figure shows, the NTCIR-5 and NTCIR-6 topic sets contain 97 topics in total, and share the 858,400 Mainichi and Yomiuri documents.

Participants will be asked to submit *exactly one* completely automatic run using the DESCRIPTION fields of the 97 topics [2]. Exactly one run because

[1] It should be noted, however, that MAP is a metric for the task of finding *all* relevant documents. For the task of finding *one* relevant document, other metrics such as Reciprocal Rank are available [8].

[2] Unlike TREC, the DESCRIPTION fields and the TITLE fields of the NTCIR CLIR test collections are similar in terms of query length and performance: The only essential difference is that the

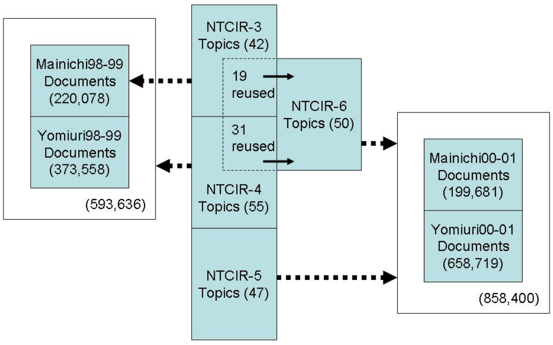
Mainichi98-99 Documents (220,078)
Yomiuri98-99 Documents (373,558)
(593,636)
NTCIR-3 Topics (42)
19 reused
31 reused
NTCIR-4 Topics (55)
NTCIR-6 Topics (50)
NTCIR-5 Topics (47)
Mainichi00-01 Documents (199,681)
Yomiuri00-01 Documents (658,719)
(858,400)

Figure 1. The NTCIR-3,4,5,6 Japanese collections.

the participants already have the *qrels* (right answers): they can compare different runs within their own group and select one as the official one for submission. (Participants will be asked *how* the submitted run has been selected, e.g., which evaluation metric was used primarily.) The run file must be in the standard TREC / NTCIR format, i.e., a trec_eval-readable format containing up to 1000 document IDs for each topic.

The above (tentative) task definition means that, for those who participated in the NTCIR-5 or NTCIR-6 Japanese document retrieval subtask, it is extremely easy to participate in USAT. They already have the document index; they already have the topics and qrels.

The key difference is that USAT will rank the participants' runs based on *Mean User Satisfaction* (MUS), rather than automatically computable metrics such as MAP. We will announce to the participants in advance as follows: "*Three assessors will each give a user satisfaction score that ranges between 0 and 1 to each of your ranked lists. Each assessor will only examine the top 10 documents of each ranked list, without referring to the qrels files, and assess the ranked list as a whole: He/She does not explicitly judge the relevance of each document.*" In short, the goal of each participant is to produce a run that human assessors *like*, rather than those that yield high evaluation metric values.

Formally, let $o(T, S)$ denote the ranked output from participating System S for Topic T. Let $us(T, S, J)$ denote the user satisfaction score given by Judge J for $o(T, S)$. This can be averaged across the three judges to alleviate the problem of inter-judge disagreement:

$$US(T,S) = \frac{\sum_J us(T,S,J)}{\sum_J} \ . \qquad (1)$$

In practice, we will employ different teams of judges for different topics since it is difficult for one person to handle all of the 97 topics. Note also that the NTCIR-5/NTCIR-6 relevance assessors and the NTCIR-7 USAT ranked output assessors are different people, although we will probably ignore this fact in our analyses.

The criterion for ranking the participants' systems, which we call Mean User Satisfaction, is just $US(T, S)$ averaged across the topic set:

$$MUS(S) = \frac{\sum_T US(T,S)}{\sum_T} \ . \qquad (2)$$

3 Assessor Effort

Each assessor will be provided with a minimal IR interface that displays the titles (and possibly keyword-in-context snippets) of the top 10 documents for given T and S. At the top of the window, the DESCRIPTION field and the NARRATIVE field of T will be shown in order to bridge the gap between the new USAT assessors and the original document relevance assessors. The assessor will be given up to 20 minutes ($1/3$ hour) to examine the top 10 documents: This should be more than enough, considering how little time people spend on looking at the first page of a Web search result, for instance. When the time is up, or when the assessor has viewed every document at least once by clicking on its title, he/she will enter a user satisfaction score: $0, 0.1, \ldots,$ or 1.

Eleven teams participated in the NTCIR-6 CLIR Japanese monolingual subtask [6]: If we have 20 participating teams and therefore 20 submitted runs at USAT, this yields $20 * 97 = 1940$ ranked lists. We need three assessors for each ranked list so a total of $3*1940$ ranked lists must be judged. Since it takes $1/3$ hour to judge one ranked list, the total time required for assessment is 1940 hours. For the NTCIR-6 CLIR releavance assessments, five Japanese assessors were employed: If we also empoly five assessors for USAT as well, the entire assessment process can be handled within $1940/5 < 400$ hours, in theory.

In addition, we can also reuse some of the runs submitted to the NTCIR-6 task. For example, Toshiba's monolingual Stage 1 run TSB-J-J-D-02 and Stage 2 run TSB-J-J-D-02-N5 [10] can be merged so that it may be treated as a single USAT run [3]. Suppose we take exactly one DESCRIPTION run from each NTCIR-6 participant: Then we have 11 additional runs. Thus, instead of 20 runs, we may have about 30 runs in total. If we have five judges, the entire assessment process can be handled in about $30/20 * 400 = 600$ hours. Budget permitting, we may double the number of assesors.

DESCRIPTIONs are sentences while the TITLEs are a list of words that lack a syntactic structure.

[3] The two Toshiba runs used exactly the same IR strategy, but other teams may have used different strategies for Stages 1 and 2.

4 Handling the Outcome

As for how to analyse the results output from this task, we are open to suggestions. As we shall mention in the next section, any participants can analyse the entire USAT results if they choose to.

Our tentative plan is to examine several existing, well-documented IR metrics such as Average Precision, Precision at 10, Reciprocal Rank, normalised Discouned Cumulative Gain (nDCG) [5] and Q-measure [7, 9]. Let $M(T, S)$ denote the value of Metric M for Topic T and System S. We will at least look at how $M(T, S)$ is correlated with $U(T, S)$: If we want to design an evaluation metric that can predict the *rankings* of systems based on user satisfaction, we can look at Spearman's or Kendall's rank correlation; If we want one that directly estimates the true absolute value of the user satisfaction score, we can use the linear correlation coefficient or Root Mean Square error. We can also examine the topic effect and the system effect on $U(T, S)$ as well as on $M(T, S)$ using statistical techniques such as analysis of variance. We may even design new evaluation metrics using regression techniques, where the useful explanatory variables may be basic statistics such as term frequency, document frequency and document length of the retrieved documents, or possibly some "meta statistics" such as the similarity of the submitted ranked lists [1]. We may also look at how $MUS(S)$ is correlated with *Mean* Average Precision and so on.

5 Tentative Schedule

If USAT is accepted as a new task for NTCIR-7, we would like to proceed as follows:

Jun 2007	USAT task specification meeting (NII, Tokyo)
Aug 2007	USAT participants registration starts
Nov 2007	USAT participants registration ends
Jan 2008	USAT runs due
Jun 2008	USAT user satisfaction scores and run files released to all participants
Oct 2008	NTCIR-7 papers due
Dec 2008	NTCIR-7 Workshop Meeting (NII, Tokyo)

The USAT Japanese test collection (derived from the NTCIR-5 and NTCIR-6 test collections), with qrels, will be released to USAT participants as soon as they register (between August and November 2007). In June 2008, the ranking of the submitted runs will be released, together with the raw user satisfaction data *plus all runs submitted to USAT*. Hence, any participant, not just the organisers, will have immediate access to all the data there is. We also plan to release all the data to non-participants on demand, probably after the NTCIR-7 Workshop Meeting in December 2008.

If we can successfully come up with new evaluation metrics based on the NTCIR-7 USAT results. the USAT task may be continued until NTCIR-8, so that the usefulness of the new metrics can be examined further.

6 Getting Involved

There are three ways to get involved in USAT:

1. Be a participant;
2. Be a task organiser[4];
3. Be a participant *and* a task organiser!

Or you can start by becoming an observer. To be an observer (and possibly a participant/organiser in the future), send an email to sakai@newswatch.co.jp and join our mailing list!

References

[1] Aslam, J. A. and Savell, R.: On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments, *ACM SIGIR 2003 Proceedings*, pp. 361-362, 2003.

[2] Buckley, C. and Voorhees. E. M.: Retrieval Evaluation with Incomplete Information, *ACM SIGIR 2004 Proceedings*, pp. 25-32, 2006.

[3] Della Mea, V. & Mizzaro, S.: Measuring Retrieval Effectiveness: A New Proposal and a First Experimental Validation. *Journal of the American Society for Information Science and Technology*, Vol. 55, No. 6, pp. 530-543, 2004.

[4] Gey, F. C., Kando, N., Lin, C.-Y. and Peters, C.: SIGIR 2006 Workshop Report: New Directions in Multilingual Information Access, *SIGIR Forum*, Vol. 40, No. 2, 2006. http://www.acm.org/sigs/sigir/forum/2006D/2006d_sigirforum_gey.pdf

[5] Kekäläinen, J.: Binary and Graded Relevance in IR evaluations - Comparison of the Effects on Ranking of IR Systems, *Information Processing and Management*, Vol. 41, pp. 1019-1033, 2005.

[6] Kishida, K. *et al.*: Overview of CLIR Task at the Sixth NTCIR Workshop, *NTCIR-6 Proceedings*, 2007.

[7] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR 2006 Proceedings*, pp. 525-532, 2006.

[4] Masao Takaku, who has already done a pilot study that is similar in spirit to USAT (Visit his EVIA 2007 poster [13] too!), has agreed to join the organising team of USAT.

[8] Sakai, T.: Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document, *AIRS 2006 Proceedings*, LNCS 4182, pp. 374-389, 2006.

[9] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, Vol. 43, Issue. 2, pp. 531-548, 2007.

[10] Sakai, T. *et al.*: Toshiba BRIDJE at NTCIR-6 CLIR: The Head/Lead Method and Graded Relevance Feedback, *NTCIR-6 Proceedings*, 2007.

[11] Sanderson, M. and Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, *ACM SIGIR 2005 Proceedings*, pp. 162-169, 2005.

[12] Soboroff, I.: A Comparison of Pooled and Sampled Relevance Judgments in the TREC 2006 Terabyte Track, *EVIA 2007 (NTCIR-6 Pre-Meeting Workshop) Proceedings*, 2007.

[13] Takaku, M., Egusa, Y., Saito, H. and Terai, H.: An Application of the NTCIR-WEB Raw-data Archive Dataset for User Experiments, *EVIA 2007 (NTCIR-6 Pre-Meeting Workshop) Proceedings*, 2007.

[14] Turpin, A. and Scholer, F.: User Performance versus Precision Measures for Simple Search Tasks, *ACM SIGIR 2006 Proceedings*, pp. 11-18, 2006.

[15] Voorhees, E. M.: The Philosophy of Information Retrieval Evaluation, *CLEF 2001 Proceedings*, LNCS 2406, pp. 355-370, 2002.

[16] Voorhees, E. M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error, *ACM SIGIR 2002 Proceedings*, pp. 316-323, 2002.

[17] Yilmaz, E. and Aslam, J. A.: Estimating Average Precision with Incomplete and Imperfect Judgments, *CIKM 2006 Proceedings*, 2006.