The Good, the Bad, the Difficult, and the Easy: Something Wrong with Information Retrieval Evaluation?

Stefano Mizzaro

Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206
Udine, Italy
mizzaro@dimi.uniud.it
http://www.dimi.uniud.it/mizzaro/

Abstract. TREC-like evaluations do not consider topic ease and difficulty. However, it seems reasonable to reward good effectiveness on difficult topics more than good effectiveness on easy topics, and to penalize bad effectiveness on easy topics more than bad effectiveness on difficult topics. This paper shows how this approach leads to evaluation results that could be more reasonable, and that are different to some extent. I provide a general analysis of this issue, propose a novel framework, and experimentally validate a part of it.

Keywords: Evaluation, TREC, topic ease and difficulty.

1 Introduction

As lecturers, when we try to assess a student's performance during an exam, we distinguish between easy and difficult questions. When we ask easy questions to our students we expect correct answers; therefore, we give a rather mild positive evaluation if the answer to an easy question is correct, and we give a rather strong negative evaluation if the answer is wrong. Conversely, when we ask difficult questions, we are quite keen to presume a wrong answer; therefore, we give a rather mild negative evaluation if the answer to a difficult question is wrong, and we give a rather strong positive evaluation if the answer is correct.

The difficulty amount of a question can be determined a priori (on the basis of lecturer's knowledge of what and how has been taught to the students) or a posteriori (e.g., by averaging, in a written exam, the answer evaluations of all the students to the same question). Probably, a mixed approach (both a priori and a posteriori) is the most common choice.

During oral examinations, when we have an idea of student's preparation (e.g., because of a previous written exam, or a term project, or after having asked the first questions), we even do something more: we ask difficult questions to good students, and we ask easy questions to bad students. This sounds quite obvious too: what's the point in asking easy questions to good students? They will almost

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 642–646, 2008.

certainly answer correctly, as expected, without providing much information about their preparation. And what's the point in asking difficult questions to bad students? They will almost certainly answer wrongly, without providing much information — and incidentally increase examiner's stress level.

Therefore we can state the following principles, as "procedures" to be followed during student's assessment:

Easy and Difficult Principle. Weight more (less) both (i) errors on easy (difficult) questions and (ii) correct answers on difficult (easy) questions.

Good and Bad Principle. On the basis of an estimate of student's preparation, ask (i) difficult questions to good students and (ii) easy questions to bad students.

I am not aware of any lecturer/teacher/examiner which would not agree with the two principles, and which would not behave accordingly, once enlightened by them.

In Information Retrieval (IR) evaluation we are not enlightened, and we do not behave like that, at least so far. In TREC-like evaluation exercises [4], all topics are equal and concur equally to determine IR system effectiveness. If a topic is "easy" (e.g., systems are highly effective on it), and an IR system performs well on that topic, the system gets a boost in its overall effectiveness which is equal to the boost it would get when performing well on a more "difficult" topic. Vice versa, if a topic is "difficult", and an IR system performs poorly on that topic, the system gets a penalty in its overall effectiveness which is equal to the penalty it would get when performing poorly on a more "easy" topic.

The only related approach is to select the difficult topics (a posteriori, on the basis of average systems effectiveness) and to include them in the Robust Track [3]. However, this is of course quite different from the two above stated principles: it would correspond to ask difficult questions only, and anyway all the difficult topics are equally difficult. Also, the effectiveness metric used in the Robust Track (i.e., the GMAP, Geometric Mean Average Precision [2]) gives more weigh to changes in the low end of the effectiveness scale, i.e., to difficult topics, but this is again limited when compared to the two above stated principles.

Furthermore, in IR evaluation we do not take into account ease and difficulty neither at the document level: given a topic, if the relevance estimation of a document by an IR system is "easy" (i.e., it is easy to determine if the document is relevant or nonrelevant — or partially relevant or whatever — to the topic) and an IR system performs well on that document, the system gets a boost in its overall effectiveness which is equal to the boost it would get when performing well on a more "difficult" document. And vice versa. Even worse, when a system is performing well (poorly), it is asked to continue to answer easy (difficult) topics and to rank easy (difficult) documents, which it will likely do with good (bad) performance.

This paper is a first attempt to address these issues. I just concentrate on the first principle at the topic level; the other issues are left for future work.

Table 1. Good, Bad, Difficult, Easy

		Effectiveness (AP)	
		Bad	**Good**
Difficulty	**Difficult**	–	++
	Easy	− −	+

2 Ease and Difficulty

A first binary view is represented in Table 1: a good effectiveness on a difficult topic should increase system effectiveness a lot (++); a good effectiveness on an easy topic should increase system effectiveness by a small amount, if any (+); a bad effectiveness on an easy topic should decrease system effectiveness a lot (− −); a bad effectiveness on a difficult topic should decrease system effectiveness by a small amount, if any (–).

Effectiveness can be defined, as usual in TREC, by means of AP (Average Precision, the standard effectiveness measure used in TREC): a high AP of a system on a topic means that the system is effective on the topic, although this neglects the ease/difficulty dimension. In a TREC-like setting, difficulty can be defined in a natural way a posteriori, as $1 - \text{AAP}$, where AAP (Average Average Precision [1]) is the average of AP values across all systems for a single topic. Hence, the difficult topics are those with a low AAP, i.e., the topics with a low average effectiveness of the systems participating in TREC. Of course this is just one among all the possible alternatives, since topic difficulty could be defined, e.g., by considering the minimum effectiveness in place of the average, or the maximum effectiveness, or by considering the best systems only, etc.

Therefore, a high AP (Average Precision, the standard effectiveness measure used in TREC) of a system on a topic could mean not only good system (high effectiveness) but also easy topic (low difficulty); conversely, low AP means bad system (low effectiveness) and/or difficulty (high difficulty).

There are several (actually, infinite) ways to turn the binary view into a continuous one. In this paper I stick with a possible choice, i.e., the function shown in Figure 1 and defined as

$$\text{NAP}(e, d) = [(1 - d) \cdot M_E + d \cdot (1 - m_D)] \cdot e^{K^{1-2d}} + d \cdot m_D .$$

This is a function from $[0, 1]^2$ into $[0, 1]$, the two variables being system effectiveness e and topic difficulty d (measured, respectively, as AP and $1 - \text{AAP}$). The result is NAP, a "normalized" version of AP values, that takes into account topic difficulty: $\text{NAP}(e, d)$ has higher values for higher e, and it has higher values, and increases more quickly, for higher d (right hand side of the figure). M_E is the maximum NAP value that can be obtained on an easy ($d = 0$, $\text{AAP} = 1$) topic. m_D is the minimum NAP value that can be obtained on a difficult ($d = 1$) topic. The model could include other 2 parameters m_E and M_D with obvious meanings, but it is natural to set $m_E = 0$ and $M_D = 1$. Also, in the figure and

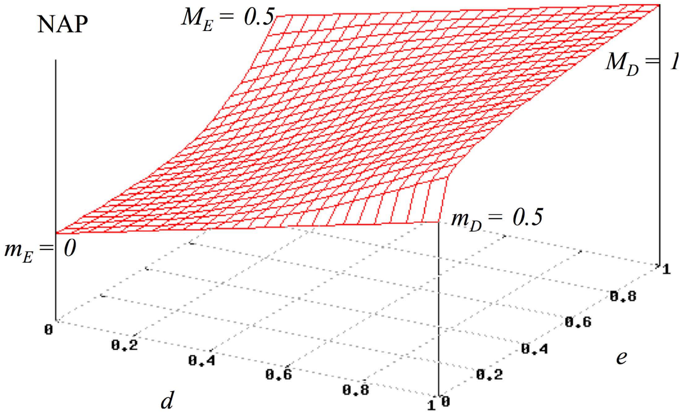

Fig. 1. The normalization function

in the following experiments, $M_E = m_D = \frac{1}{2}$. $K \geq 1$ allows different curvatures (in figure, $K = 4$; in the following, $K = 100$; higher K values lead to stronger normalizations, but for lack of space the role of K is not discussed here).

The proposed function is just one among infinite possible choices: Table 1 just sets some constraints on the four corners of Figure 1 (d and $e \in \{0, 1\}$); the chosen parameters values m_D, m_E, M_E, and M_D, satisfy these constraints, but of course their values could be different; and the interpolation of the four corners could be done in infinite ways. The study of variants is left as future work.

3 Experiments and Results

Averaging across topics the NAP values obtained as above described, we obtain a new measure of retrieval effectiveness, that I name NMAP, for Normalized MAP (Mean Average Precision). We can then compare retrieval effectiveness as measured by MAP and NMAP. I use data from TREC 8 (129 systems, 50 topics).

Figure 2 shows the differences in ranking of the 129 systems participating in TREC when their effectiveness is measured by MAP and NMAP. It is clear from the scatterplot that the two rankings are quite different, although related (Kendall's tau correlation is 0.87, linear correlation is 0.92). This means that by using NMAP instead of MAP one would get different rankings of the systems participating in TREC. In other words, what is generally considered an improved version of a system (a version with a higher MAP) would often turn out to be not an improvement at all when using NMAP, which is based on the reasonable assumptions sketched in Section 1. As the figure shows, MAP and NMAP do quite agree on the best systems, those in the first 20 positions or so, with very few exception (see the left hand side of the figure). However, the agreement decreases after the 20th system, with strong disagreement for a dozen of systems (the dots that stand out).

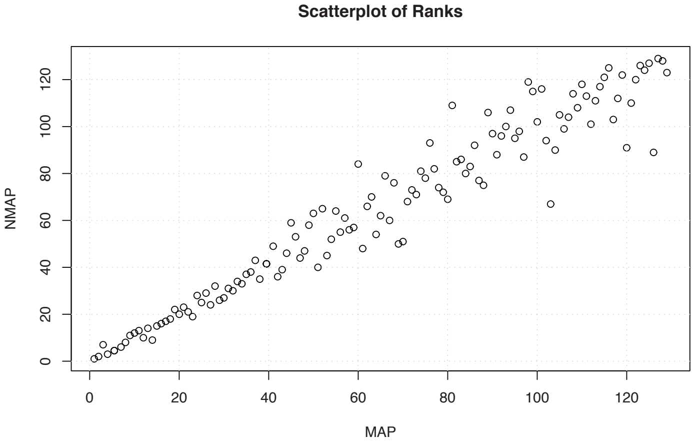

Fig. 2. Differences in systems rankings

4 Conclusions and Future Work

These preliminary experiments do indeed hint that if we followed the first principle stated in Section 1, TREC results could be somewhat different (in terms of both system ranking and absolute effectiveness values): we might be evaluating TREC systems in a wrong way.

This paper can be seen as a research agenda, since further work is needed to confirm these results, on several aspects. The normalization function could be improved (e.g., it could be rewritten with a GMAP [2] flavor, exploiting logarithms). It will be interesting to see what happens when the second principle is considered as well, since this might lead to reduce the number of topics used in TREC-like evaluations, and when the same analysis is extended to the document level. From a different point of view, NMAP is a new metric for retrieval effectiveness; it will be interesting to study its relationships with other metrics (like GMAP, which seems to be a special case of NMAP), its general properties (e.g., stability), and its relationship with user satisfaction (by means of user studies).

References

1. Mizzaro, S., Robertson, S.: HITS Hits TREC - Exploring IR Evaluation Results with Network Analysis. In: 30th SIGIR, pp. 479–486 (2007)
2. Robertson, S.: On GMAP – and other transformations. In: 13th CIKM, pp. 78–83 (2006)
3. Voorhees, E.M.: Overview of the TREC 2005 Robust Retrieval Track. In: TREC 2005 Proceedings (2005)
4. Voorhees, E.M., Harman, D.K.: TREC — Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (2005)