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 — Italy

Easy and difficult questions

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 we expect correct answers ⇒
  - rather mild positive evaluation if the answer to an easy question is correct
  - rather strong negative evaluation if the answer is wrong
- Conversely, when we ask **difficult** questions, we rather presume a wrong answer ⇒
  - rather mild negative evaluation if the answer to a difficult question is wrong
  - rather strong positive evaluation if the answer is correct

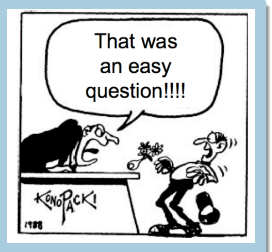

Figure: Ever feel like this?

Good and bad students

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**:

- What's the point in asking easy questions to **good** students? They will almost certainly answer correctly, as expected, without providing much information about their preparation.
- 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.

(http://www.listen-project.de/garfield/index.php?date=07.02.2008)

Figure: Or like this?!

Student assessment: Two principles

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.

IR evaluation

- TREC-like experiments
- All topics are equal (Yes, we have the Robust track — but it's limited.)
- All documents are equal
- When a system is good (bad), it continues to work on easy (difficult) topics and on easy (difficult) documents
- System effectiveness on a topic = AP (Average Precision)
- Overall system effectiveness = MAP (Mean Average Precision)
- **Topic ease = AAP (Average Average Precision)**

	t_1	$\cdots$	t_n	MAP
s_1	$AP(s_1, t_1)$	$\cdots$	$AP(s_1, t_n)$	$MAP(s_1)$
$\vdots$		$\ddots$		$\vdots$
s_m	$AP(s_m, t_1)$	$\cdots$	$AP(s_m, t_n)$	$MAP(s_m)$
AAP	$AAP(t_1)$	$\cdots$	$AAP(t_n)$	

Table: AP, MAP, and AAP

Easy and Difficult Principle — IR version

- Weight more both: (i) low AP on easy (high AAP) topics and (ii) high AP on difficult (low AAP) topics.
- Weight less both: (i) low AP on difficult (low AAP) topics and (ii) high AP on easy (high AAP) topics.

Normalized MAP: NMAP

Binary view:

- 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 (AP)	
		Bad	Good
Difficulty (AAP)	Difficult	−	+++
	Easy	−−−	+

Table: Good, Bad, Difficult, Easy

Continuous view:

- **NAP** = Normalized Average Precision; **NMAP** = Mean NAP
- Four corners → four entries in the table above, with (arbitrary) choice of values
- Non linearity:
  - On a difficult topic, a small increase in (low) AP is immediately rewarded
  - On an easy topic, a small decrease in (high) AP is immediately rewarded

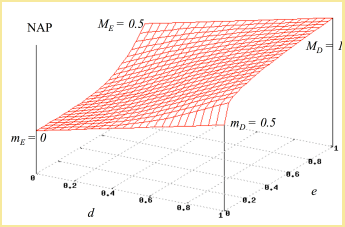

Figure: Normalization function (difficulty $d = 1 - AAP$; effectiveness $e = AP$)

Results: MAP vs. NMAP

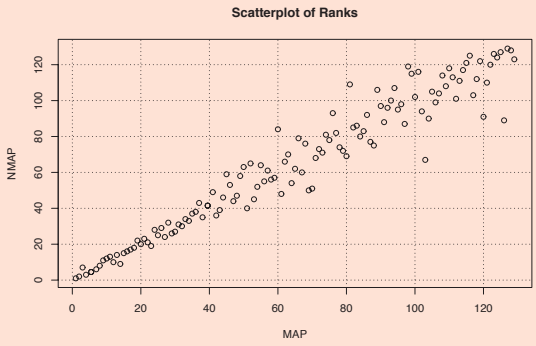

Figure: Differences in systems rankings

- using NMAP instead of MAP ⇒ different rankings of the systems participating in TREC
- Kendall's tau correlation is 0.87
- linear correlation is 0.92
- 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 above.
- MAP and NMAP do quite agree on the best (∼ 20) 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).

Conclusions & future

Main result:

- If we followed the "Easy and Difficult Principle — IR version" stated above, 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.

Future:

- Improve the normalization function
- Consider the second "Good and Bad Principle" (reduce the number of topics in TREC?!)
- Work at the document (not only topic) level
- Compare NMAP with other metrics (e.g., GMAP)

ECIR 2008 Proceedings: pp. 642–646

http://www.dimi.uniud.it/mizzaro/ mizzaro@dimi.uniud.it