Fading away: Dilution and user behaviour

Paul Thomas (CSIRO & ANU), Falk Scholer (RMIT University), Alistair Moffat (University of Melbourne)

www.csiro.au

Faced with a poorly-performing search system, users might react in different ways: for example they might issue more queries, read more deeply, or take more time.

We controlled system quality and task difficulty to measure these reactions. Although users could tell the difference between "good" and "bad" systems, **their behaviour was very similar**. This means we need to be careful interpreting user behaviour to infer system quality.

Tasks, systems, and participants

Our experiment used:

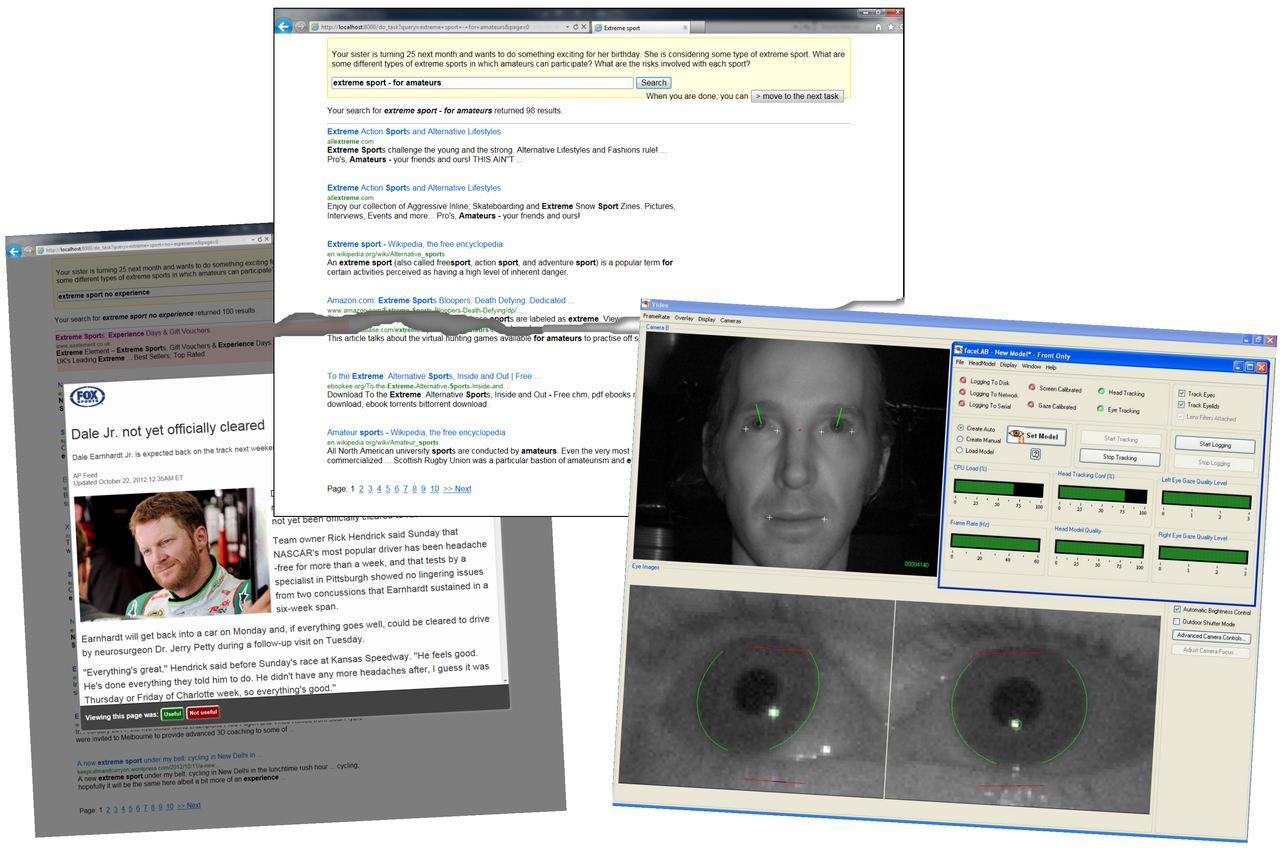
- Six **tasks of controlled difficulty**, from Wu et al, across three types: "remember", "understand", and "analyse".
- And two **search systems of controlled quality**. The "full" system presented results from Yahoo!. The "diluted" system had bad, but plausible, results at ranks 1, 3, 5, ... Users were given the first query but allowed to issue new queries, move to later pages, or view results in a pop-up window.

34 participants completed each of the six tasks, three each ("remember", "understand", "analyse") on each of the systems ("good", "diluted").

Measures

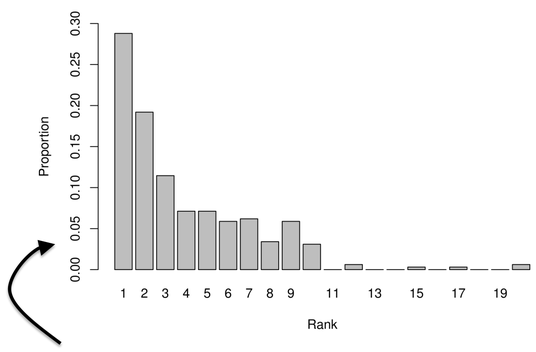
We measured several aspects of **user behaviour**:

- Time,
- Clicks,
- Queries, and
- Clicks,
- Scrolls,
- Gaze.

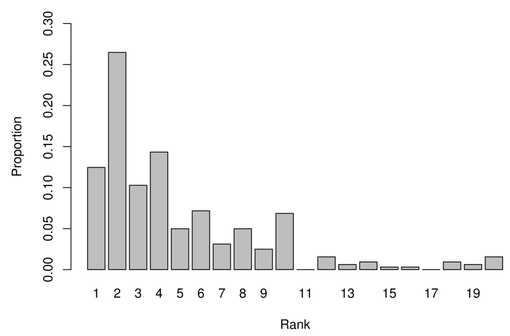


Results

Our users **could tell the difference** between the good and bad results, and did not often click on the bad ones:

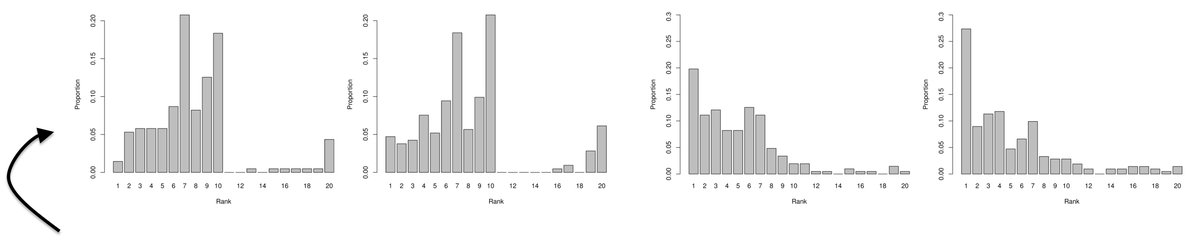


Full: all results are normal



Diluted: every odd-numbered result is bad

But other aspects of behaviour were very similar indeed between the two systems, including **views**:



Deepest rank viewed

Final rank viewed

And **queries** (median 1 query for full/2 for diluted, difference is not significant);

And **everything else** (rate of saves 79%/75% ; reading time per snippet 0.59s/0.59s; pagination 7%/10%; reported satisfaction 4/4 out of 5; reported difficulty 2/2 out of 5; and so on).

So what?

Our results are consistent with observations that small differences in e.g. MAP don't make a real difference in user performance (Turpin & Scholer).

We need to be careful using changes in behaviour to evaluate search systems. It's possible to have significantly lower system performance with no noticable changes in user behaviour at all.

FOR FURTHER INFORMATION

Paul Thomas

e paul.thomas@csiro.au

w es.csiro.au

REFERENCES

Turpin and Scholer. User performance versus precision measures for simple web search tasks. In *Proc. SIGIR*, 2006.

Wu, Kelly, Edwards, and Arguello. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proc. IIiX*, 2012.