Causality, prediction and improvements that (don't) add up

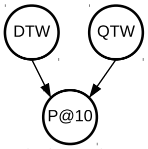
– Position Paper–

Norbert Fuhr
University of Duisburg-Essen
Germany
norbert.fuhr@uni-due.de

A recent Dagstuhl workshop focused on the problem of performance prediction in IR and related fields and formulated a mid-term research program [1]. As a core problem, the development of appropriate models for predicting performance was identified. In this paper, we want to show the relationship of this issue to causal inference [2], and how methods from this area may be helpful in our field.

As a starting example, let us assume that we want to model the effect of different types of document term weighting (DTW) – binary vs. tf – and query term weighting (QTW) – binary vs idf – methods on retrieval quality, where we model the latter as P@10, which can also be regarded as the probability that a random document from the top 10 is relevant. This relationship can be modeled as a Bayesian inference network. Since the variables DTW and QTW are obviously not independent factors wrt. performance, the arrows point towards the P@10 node. (In contrast, in the binary independence retrieval model, the arrows would point from the P@10/relevance node towards the different terms).

DTW	QTW	P@10
bin	bin	0.1
bin	idf	0.5
tf	bin	0.2
tf	idf	0.7



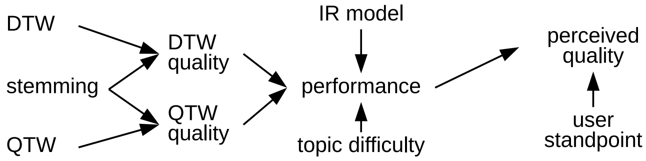
When regarding causality and the relationship of the three variables in the examples, we would have the same graph, since performance is obviously determined by both DTW and QTW. Moreover, in this example the two factors are not additive, which makes things more complex.

In [3], an ANOVA model for the influence of topic difficulty and IR system on resulting performance is investigated, where the latter is subdivided into stemming and IR model. ANOVA is based on the assumption of additivity, but the results show large residuals, so it is uncertain if additive cumulation of the different effects is the most appropriate model.

[4] investigated a broad variety of factors for improving retrieval quality, which had been proposed by various authors over a decade; each of these methods in isolation showed some improvements (although often over questionable baselines). However, when multiple factors were combined, their improvements did not add up — in most cases they were not able to yield any higher quality. We assume that the major reason for this outcome is the problem that there are strong dependencies between the different methods, which were never investigated.

As a simple causal model for retrieval quality, regard the following figure (error terms were omitted for simplicity). Here we have assumed that besides DTW and QTW methods, we also may regard different types of stemming, which influences the quality of the weights of both types. For a causal model, we need to be able to determine the values of the variables "DTW quality" and "QTW quality". For the former, we could e.g. regard the correlation of these weights with relevance (for a given set of queries) in the documents – for the latter, we might need a more sophisticated method. The retrieval model finally specifies how the two types of weights are combined with each other; the resulting performance naturally also depends on topic difficulty. In a further step, we can model the user's perception (e.g. the stopping behavior), which transforms the system-oriented performance into the perceived quality.

Using such a causal model, it is possible to investigate the influence of the different valiables on the final quality, and we are also able to regard the intermediate steps.



As can be seen from this example, causal methods may be a helpful tool for performance prediction. Besides a clear notion of the variables affecting performance (and their interdependencies), we also need observable intermediate variables, that help us in the analysis.

1. REFERENCES

[1] Nicola Ferro et al. The Dagstuhl Perspectives Workshop on performance modeling and prediction. *SIGIR Forum*, 52(1):91–101, 2018.

[2] Dana Pearl, Judea amd Mackenzie. *The Book of Why. The New Science of Cause and Effect.* Basic Books, 2018.

[3] Nicola Ferro and Gianmaria Silvello. A general linear mixed models approach to study system component effects. In *Proc. SIGIR*, 2016

[4] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In Proc. *CIKM*, 2009.