Less is Less: When are Snippets Insufficient for Human vs Machine Relevance Estimation?

Gabriella Kazai, Bhaskar Mitra, Anlei Dong, Nick Craswell, and Linjun Yang

Microsoft

Impact of body text on relevance assessments
- Body text helped in 48% cases.
- For 28%, seeing the body lead to revised label.
- Body text helped predictably poor performing, long, not-navigational, and question type queries (p<0.01).

Impact of seeing body text on neural ranker performance
- Body text helps ranker: UTSB model outperforms UTS (∆RBP>0).
- Benefit more evident at the top ranks (∆RBP@3>∆RBP@10).
- Body text can degrade performance for some queries.
- Improved queries are long, not-navigational and of question type, while degraded queries are short, head and navigational, and the documents long (p<0.01).

When does body impact human and machine relevance estimates
- Trained two EBM regression models with query/document properties as features and ΔLabel and ΔRank as targets.
  - Only common feature is body length among top 5 features that explain ΔLabel and ΔRank, resp., see Table 4.
  - Humans and machine react to body text for different query/document types.

Conclusions
- Studied when human and machine assessors benefit from the full text of the document to estimate its relevance.
- Both humans and BERT style models benefit from the body text in similar cases (for long, not navigational, tail and question type queries), but full text impacts their relevance estimations in very different ways.
- The BERT model’s performance improves or degrades with the full text depending on query property. E.g., performance degrades for navigational queries (∆RBP@3 of -1.07).
- Different types of queries (e.g., head v tail) require models to be optimized differently.