Use Cases, experiments and other requirements for STELLA II
- ftpov of ZBW / EconBiz

Timo Borst / Arben Hajra / Kim Plassmeier
ZBW – Leibniz Information Centre for Economics

‘Instructions’

- What is your envisioned use case? What kinds of experiments do you want to conduct, for instance, recommendations or document rankings?
- Do you have any experiences with user experiments, for instance, from the analysis of log files?
- What kinds of insights do you expect, and how can the online experiments with STELLA meet your (business) goals in the long run? What can we do to facilitate these goals, and what are the requirements for the STELLA framework?

Our (envisioned) use cases

- Both (re-)ranking and recommending of scientific publications from economics
- Target groups: primarily researchers / scientists / students in economics as information portal users
- LibRank model evaluation:
  - (mainly) integration of popularity data (clicks, citations, etc.)
  - Non-linear normalization scheme + prioritized aggregation
  - Mixed results from Cranfield-type evaluation (~10 experts, ~50 students)
  - Top ranked model in TREC OpenSearch 2016/17 (Schaer et al.) (but not stat. signif.)
- Integration of bibliometric indicators (not only traditional ones) as potential key research use cases
- General ability to run online user experiments (to support day-to-day operations, e.g., tuning weights, testing specific boosts, etc.)

Our user experiments so far

- LibRank relevance judgements and (re-)rankings
- SINIR simulations of user interaction behavior
- Questionnaires for exploring user expectations and interviews with researchers from economics
- Recommendation of related literature based on controlled vocabulary and/or machine learning based algorithms
- ‚labor-kind‘ of usability studies (with a focus on navigation and wording within the website)
- Log file analyses

What kind of insights do we expect?

- Does the STELLA framework really support us in deciding between variants of ranking (or even document presentation?), so A/B testing becomes more lean or even obsolete?
- What does it actually mean to setup and conduct an (online) experiment by ourselves?
- Are experiments or results with other sites transferable to our site?
- What happens after a (successful) experiment, e.g. a focused A/B test to be run afterwards?
- Does interleaving within online experiments have any side effects on a site’s performance, e.g. in case of more thorough calculations? Could these effects be limited or controlled?
- What does it mean to provide or contribute our own data, e.g. with respect to citations?

Further questions / backup

- Possibility to use session history/user data/… in models?
- Possibility to opt-out on insufficient data coverage or other data quality issues (e.g., only sparse citation data available for current query)? Just returning an empty list?
- Different experiments querying the same prod index just with different parameters (complementary to the common index)?
- Anything specific for models that only affect a few result lists?
- How does pagination work?
- Interoperability with other logging frameworks?

Use case: LibRank Ranking Model Evaluation

- (Mainly) Integration of popularity data (clicks, citations, etc.)
- Two main aspects:
  - Non-linear normalization scheme (based on Characteristic Scores and Scales method)
  - Prioritized aggregation (aka. dynamic weights)
- Mixed results from Cranfield-type evaluation (~10 experts, ~50 students)
- Top ranked model in TREC OpenSearch 2016/17 (Schaer et al.) (but not statistically significant)


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
BIYAN
SHUKRIA
TINGKI
YAQHANYELAY
TASHAKKUR ATU
SUKSAMA
EKHMET
GRAZIE
MEHRBANI
MAAKE
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
PALDIES
BOLZÏN
MERCI