FOLTR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Original | US: do findings generalise to other datasets? | US: do the findings generalise to different number of clients | US: do the findings generalise to typical OLTR evaluation setup? | US: do the findings generalise to typical OLTR evaluation setup, with MRR as evaluation? |
| Measure: MRR on train / exp MRR on test | Measure: MRR on train / exp MRR on test | Measure: MRR on train / exp MRR on test | Measure: nDCG | Measure: MRR |
| Signal: MRR | Signal: MRR | Signal: MRR | Signal: nDCG | Signal: MRR |
| Rel.Jud: Click | Rel.Jud.: Click | Rel.Jud.: Click | Rel.Jud.: labels | Rel.Jud.: labels |
| Collection: MQ | Collection: MQ, MSLR, Yahoo | Collection: MQ, MSLR, Yahoo | Collection: MQ, MSLR, Yahoo | Collection: MQ, MSLR, Yahoo? |
| Comparison | LTR: MSE, SVMRank  OLTR: PDGD | LTR: MSE, SVMRank  OLTR: PDGD | LTR: MSE, SVMRank  OLTR: PDGD | LTR: MSE, SVMRank  OLTR: PDGD |
| Privacy settings | 1, 0.5, 0.25, 0.15 | 1, 0.5, 0.25, 0.15 | 1, 0.5 | 1, 0.5 |
| Findings |  |  | FOLTR sucks – is it because of nDCG or of labels? |  |

Experiments to Run:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| dataset | client | privacy | algorithm | | conclusion |
| MQ2007 | 2000 | 1.0 | FOLtR-original | PDGD | PDGD better |
| 2000 | 0.9 | FOLtR-original |  |  |
| 2000 | 0.5 | FOLtR-original |  |  |
| 1000 | 1.0 | FOLtR-original |  |  |
| 1000 | 0.9 | FOLtR-original |  |  |
| 50 | 1.0 | FOLtR-original | PDGD |  |
| 50 | 0.9 | FOLtR-original |  |  |
| 50 | 0.5 | FOLtR-original |  |  |
| 2000 | 1.0 | Ndcg |  |  |
| 50 | 1.0 | Ndcg-todo |  |  |
| MSLR10K | 2000 | 1.0 | FOLtR-original | PDGD | PDGD better |
| 2000 | 0.9 | FOLtR-original |  |  |
| 2000 | 0.5 | FOLtR-original |  |  |
| 1000 | 1.0 | FOLtR-original-todo |  |  |
| 50 | 1.0 | FOLtR-original-todo |  |  |
|  |  |  |  |  |
| 2000 | 1.0 | Ndcg-todo-rerun |  |  |
|  |  |  |  |  |  |
| Yahoo | 2000 | 0.9 | FOLtR-original |  |  |
|  |  |  |  |  |  |