FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Back to the sketch-board: Integrating keyword search, semantics, and information retrieval

Joel Azzopardi[1], Fabio Benedetti[2], Francesco Guerra[2], and Mihai Lupu[3]

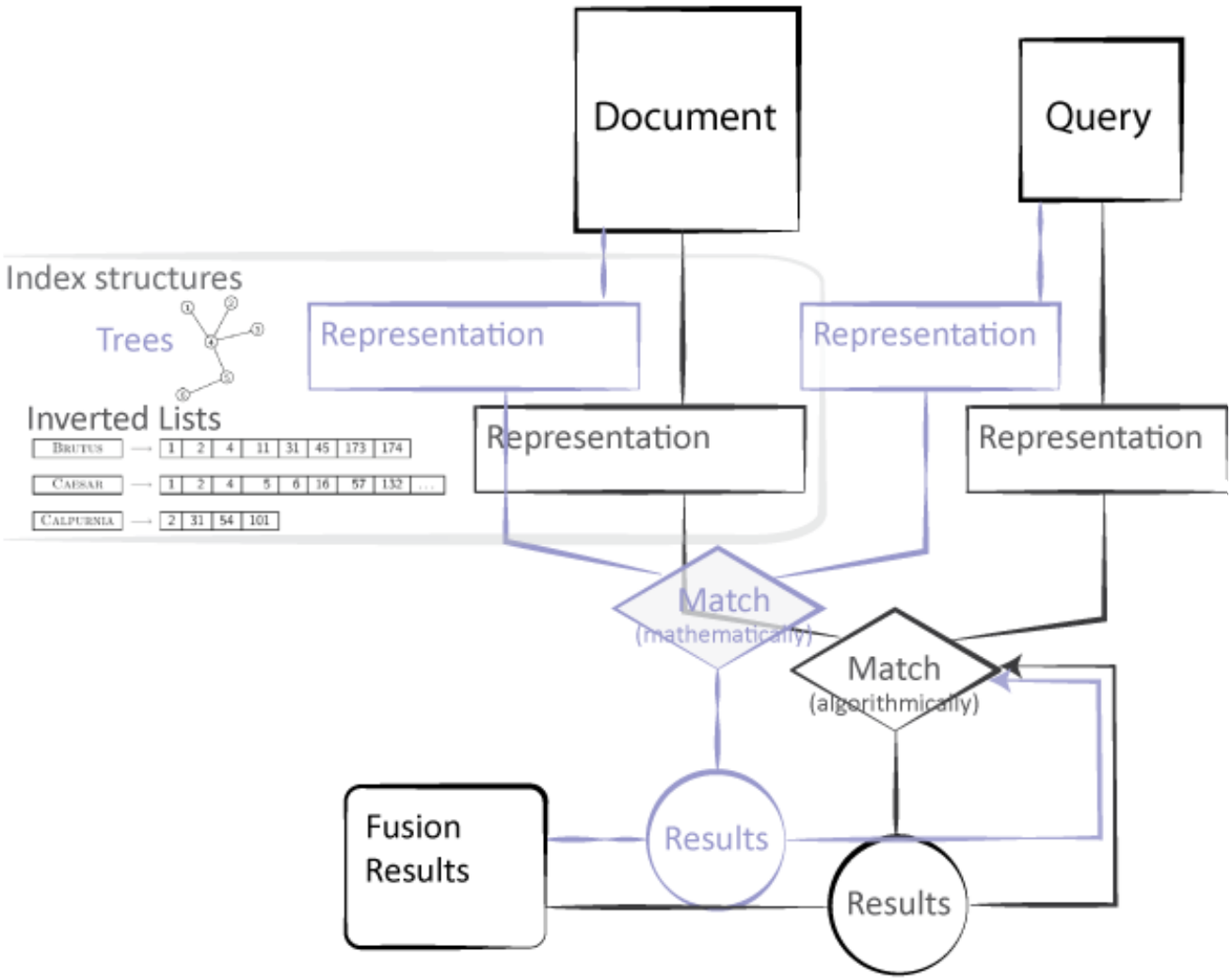
[1] University of Malta joel.azzopardi@um.edu.mt
[2] Universita di Modena e Reggio Emilia firstname.lastname@unimore.it
[3] TU Wien mihai.lupu@tuwien.ac.at

2nd International keystone Conference / IKC 2016 / Cluj-Napoca Romania, 8-9 September 2016

the sketch-board



the sketch-board



Start from existing work [KE4IR, Corcoglioniti et al. 2016]

1. experimenting new semantic representations of the data;
2. experimenting different measures for computing the closeness of documents and queries

Contributions of this paper

- we reproduce the work in KE4IR;
- we extend the work by introducing new semantic representations of data and queries;
- we change the scoring function from the tf-idf to the BM25 and BM25 variant [Lipani et al. 2016] .

1. new semantic representations

- started from a subset of the layers analyzed in KE4IR
  - only classes and entities referenced in the data
    - hypothesis: reduce the noise generated by spurious information
- extend this set in two ways:
  1. adding external classes and entities via PIKES
     - *enriched* set
  2. refine and extent annotations using DBpedia
     - use the textual description in the DBpedia *abstract* field
     - apply AlchemyAPI to it to extract additional entities.

- bm25

$$S(q,d) = \sum_{t \in T_d \cap T_q} \frac{(k_3+1)tf_q}{k3+tf_q} \frac{(k1+1)\overline{tf_d}}{k1+\overline{tf_d}} \log \frac{|D|+0.5}{df_t+0.5}$$

$$\overline{tf_d} = \frac{tf_d}{B} \qquad B = (1-b)+b\frac{L_d}{avgdl}$$

- bm25 variant

$$B_{\mathrm{VA}} = \frac{avgtf_d}{mavgtf^2} + (1 - mavgtf^{-1})\frac{L_d}{avgdl}$$

$$avgtf_d = \frac{1}{|T_d|}\sum_{t \in T_d} tf_d(t) = \frac{L_d}{|T_d|} \quad \text{and} \quad mavgtf = \frac{1}{|D|}\sum_{d \in D} avgtf_d$$

bm25 variant [Lipani et al 2016]

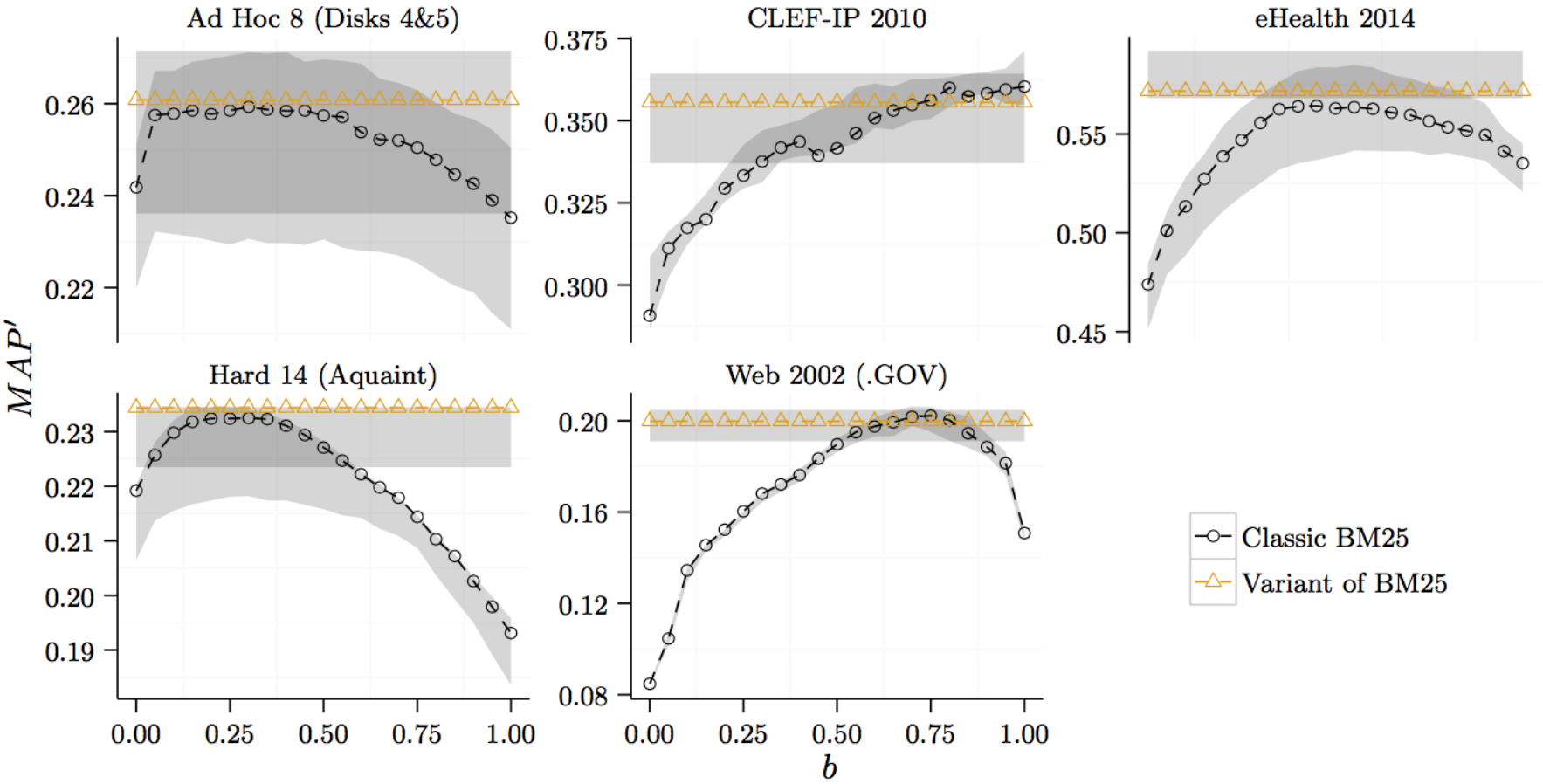


Figure 1: Performance sensitivity of the classic BM25 (CL) and our variant (VA) across all the test collections, with standard $k_1 = 1.2$. The gray area represents the range of values obtainable varying k_1 in the range $[0.5, 2.5]$.

combining terms and concepts

- Probabilistic Relevance Framework

$$d = (\overbrace{tf(t_1), tf(t_2), ..., tf(t_{|T|})}^{\text{term frequencies}}, \overbrace{ef(e_1), ef(e_2), ..., ef(e_{|E|})}^{\text{concept frequencies}})$$

- direct application not possible
  - terms and concepts do not share the same probability space
- calculated a separate $S_E(q,d)$ score

$$S_E(q,d) = \sum_{e \in E_d \cap E_q} \frac{(k_3+1)ef_q}{k3+ef_q} \frac{(k1+1)\overline{ef}_d}{k1+\overline{ef}_d} \log \frac{|D|+0.5}{df_e+0.5}$$

where

$$\overline{ef}_d = \frac{ef_d}{B_e} \qquad B_e = (1-b)+b\frac{L_d^e}{avgdl_e}$$

$$B^e_{\text{VA}} = \frac{avgef_d}{mavgef^2} + (1 - mavgef^{-1})\frac{L_d^e}{avgdl_e}$$

where

$$avgef_d = \frac{1}{|E_d|}\sum_{e \in E_d} ef_d(e) = \frac{L_d^e}{|E_d|} \quad \text{and} \quad mavgef = \frac{1}{|D|}\sum_{d \in D} avgef_d$$

combining terms and concepts

- Probabilistic Relevance Framework

$$d = (\overbrace{tf(t_1), tf(t_2), ..., tf(t_{|T|})}^{\text{term frequencies}}, \overbrace{ef(e_1), ef(e_2), ..., ef(e_{|E|})}^{\text{concept frequencies}})$$

- direct application not possible
  - terms and concepts do not share the same probability space
- calculated a separate $S_E(q,d)$ score
- combine the two

$$\mathbf{S}(q,d) = S(q,d) + \lambda S_E(q,d)$$

Experiments

1. Using terms alone comparing traditional BM25 (standard B) with the variation BVA, as well as the baseline in KE4IR;
2. Using terms (as in 1 above) after applying filtering based on concepts;
3. Combining ranking of terms and concepts; and
4. Combining ranking of terms and concepts as in 3 after applying filtering based on concepts.

Dataset

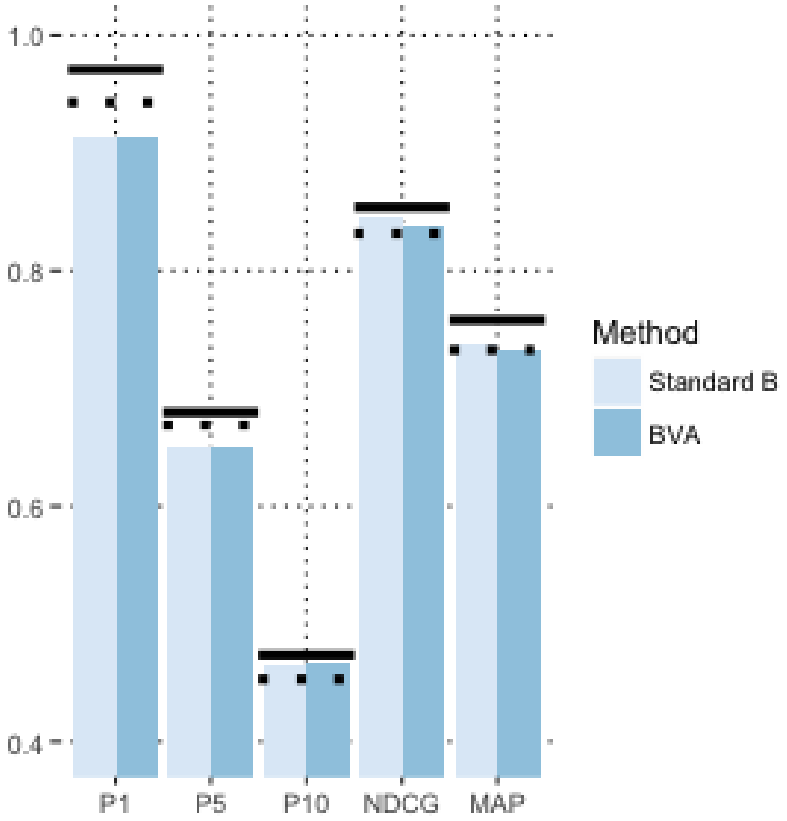
331 articles from the yovisto blog.

570 words on average

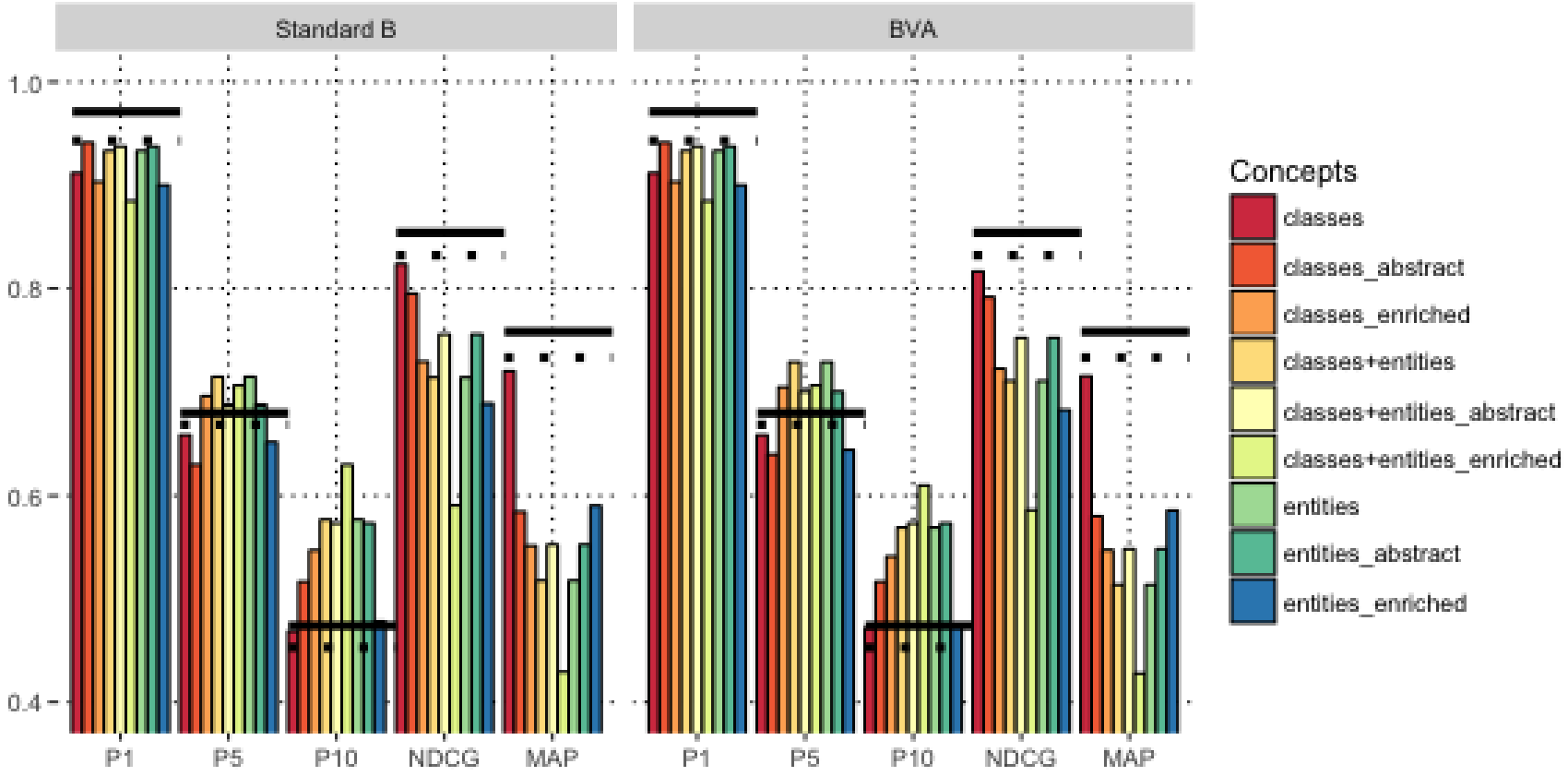
83 annotations per article, on average

35 queries inspired by search log, manually annotated

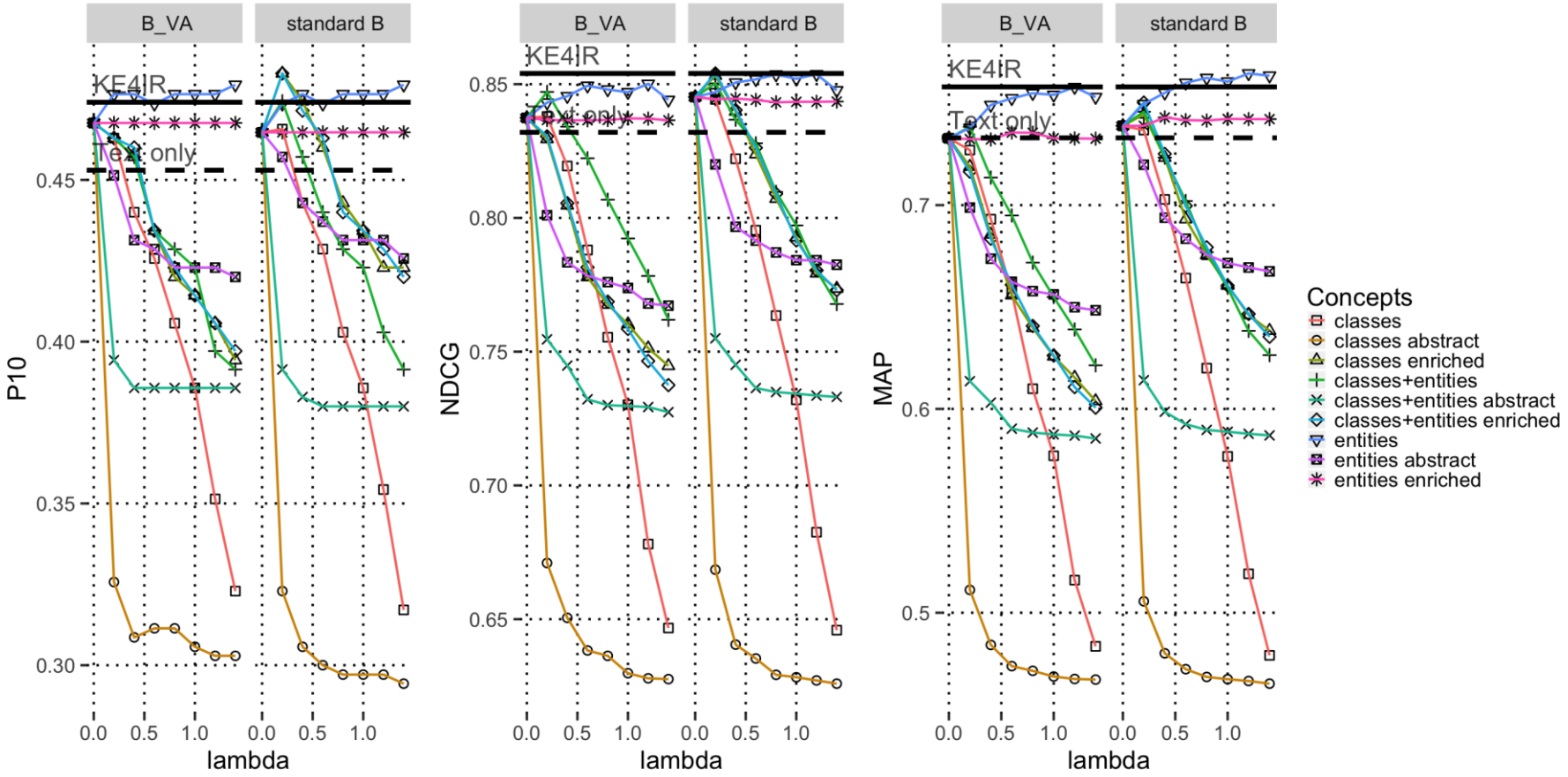
- Classic BM25 params
  - k1 = 1.2
  - k3 = 0
  - b = 0.75



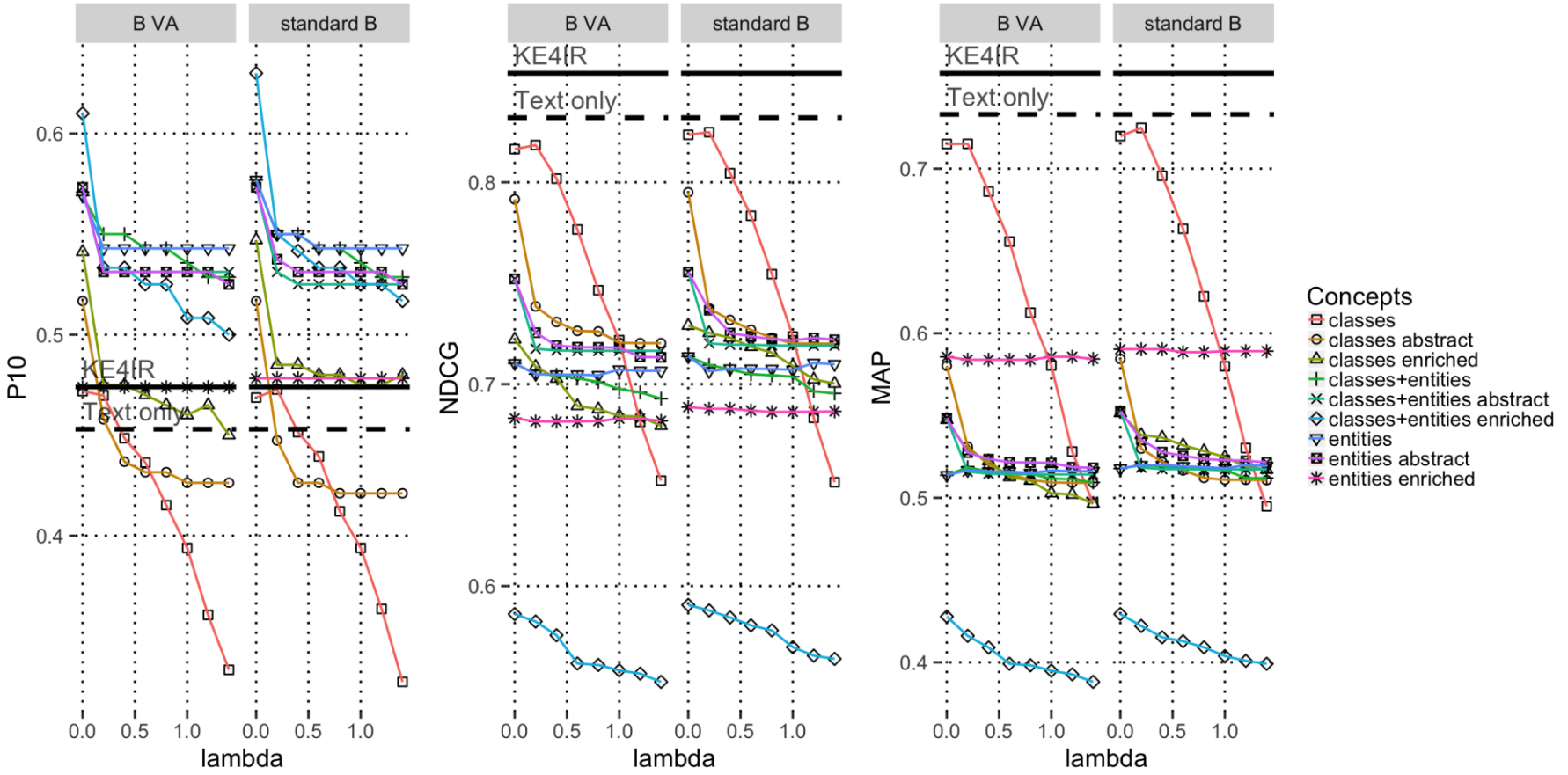
Retrieval using terms and filter on concepts



Retrieval using combined ranking of terms and concepts



Retrieval using combined ranking of terms and concepts, and filter on concepts



Observations

- Best results obtained on P@5 and P@10, improving the current state of the art on the provided test collection.
- By considering the top-heavy metrics (P@1 and MAP), the experiments show that it is extremely difficult to improve on the existing results.
- The increased performance in precision obtained by our technique does not correspond to an increase in the NDCG and MAP scores, thus meaning that a larger number of correct documents is associated to a worst ranking of them.
- The main benefit from the adoption of concepts is the filtering of the documents. Results show that in most cases they introduce more noise than utility into the ranking.
- Due to the small dataset and number of queries evaluated, the result cannot be generalized out of this domain.
- In this particular domain, the variation of BM25 introduced does not improve the scores.

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Thank you

Back to the sketch-board: Integrating keyword search, semantics, and information retrieval

Joel Azzopardi[1], Fabio Benedetti[2], Francesco Guerra[2], and Mihai Lupu[3]

[1] University of Malta joel.azzopardi@um.edu.mt
[2] Universita di Modena e Reggio Emilia firstname.lastname@unimore.it
[3] TU Wien mihai.lupu@tuwien.ac.at

2nd International keystone Conference **/ IKC 2016** / Cluj-Napoca Romania, 8-9 September 2016