Notes on Topic-sensitive PageRank

R. Basili

a.a. 2013-14

The PageRank main workflow

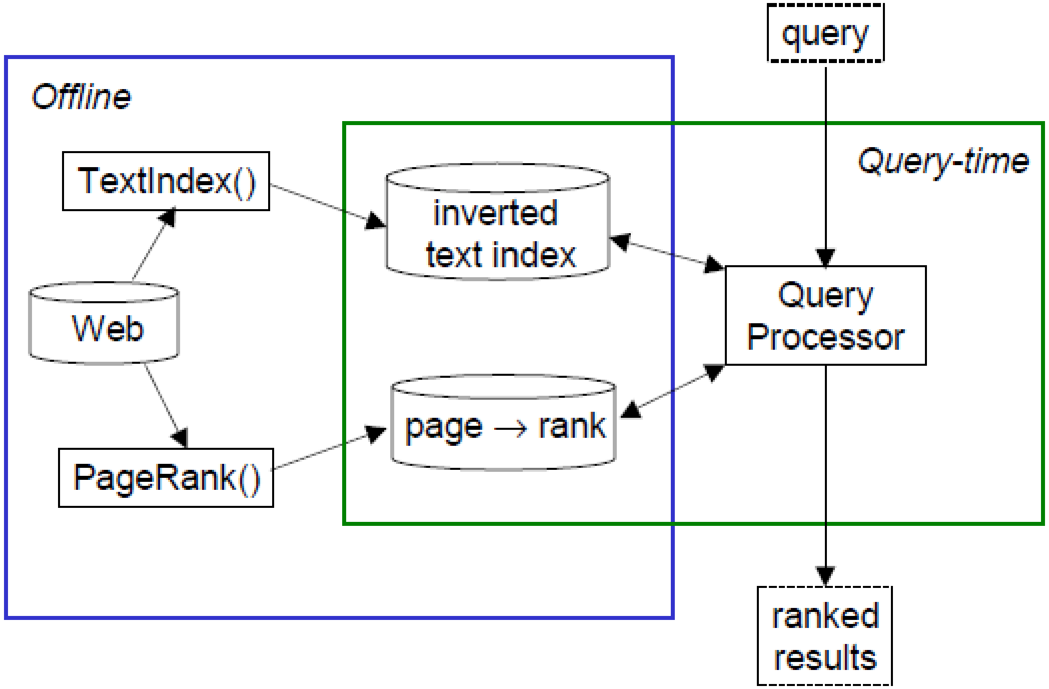


Fig. 1. Simplified diagram illustrating a simple search engine utilizing the standard PageRank scheme.

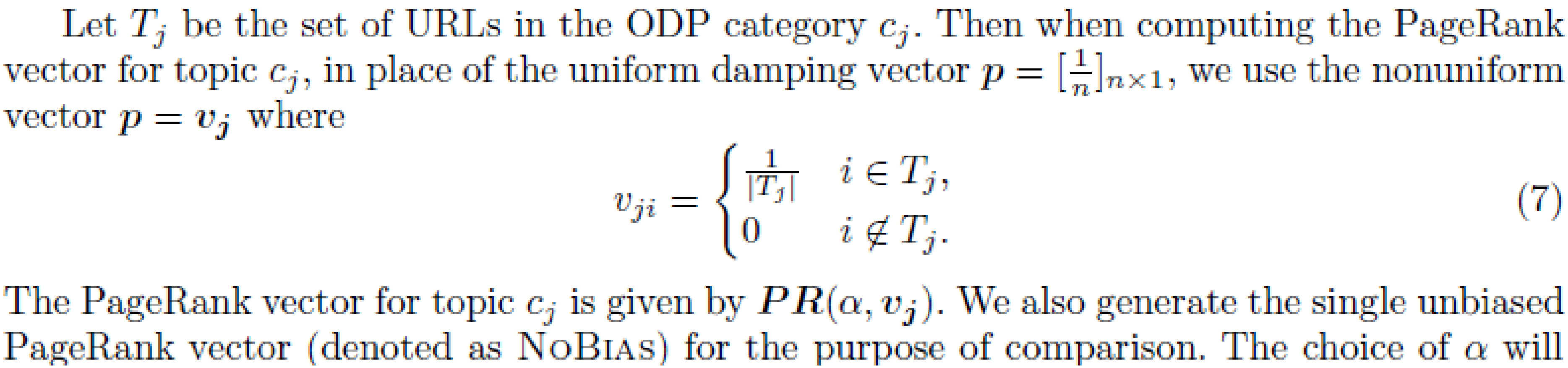
Personalizing PageRank

- IDEA: determine the topics most closely associated with the query, and use the appropriate topic-sensitive PageRank vectors for ranking the documents satisfying the query.
- This ensures that the “importance" scores reflect a preference for the link structure of pages that have some bearing on the query.
- As with ordinary PageRank, the topic-sensitive PageRank score can be used as part of a scoring function that takes into account other IR-based scores.

Original PageRank

$$\begin{aligned} Rank &= M' \times Rank \\ &= (1-\alpha)(M+D) \times Rank + \alpha p \end{aligned}$$

Topic-sensitive PageRank

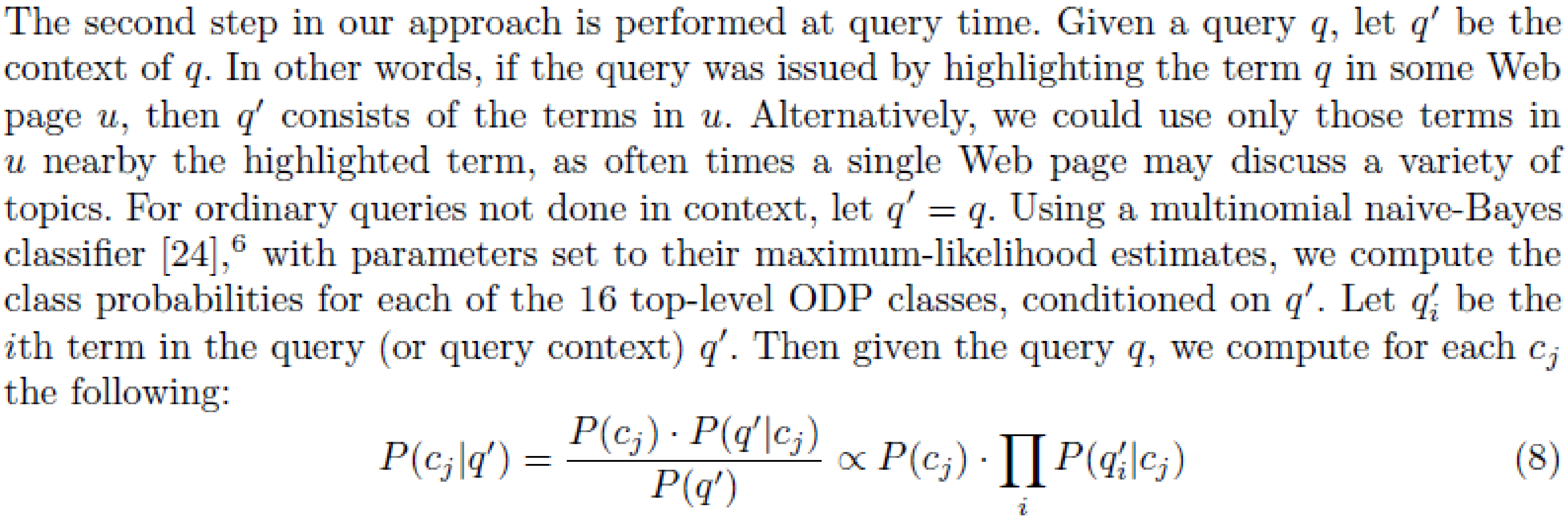


Let T_j be the set of URLs in the ODP category c_j. Then when computing the PageRank vector for topic c_j, in place of the uniform damping vector $\boldsymbol{p} = [\frac{1}{n}]_{n\times 1}$, we use the nonuniform vector $\boldsymbol{p} = \boldsymbol{v_j}$ where

$$v_{ji} = \begin{cases} \frac{1}{|T_j|} & i \in T_j, \\ 0 & i \notin T_j. \end{cases} \tag{7}$$

The PageRank vector for topic c_j is given by $\boldsymbol{PR}(\alpha, \boldsymbol{v_j})$. We also generate the single unbiased PageRank vector (denoted as NoBias) for the purpose of comparison. The choice of α will

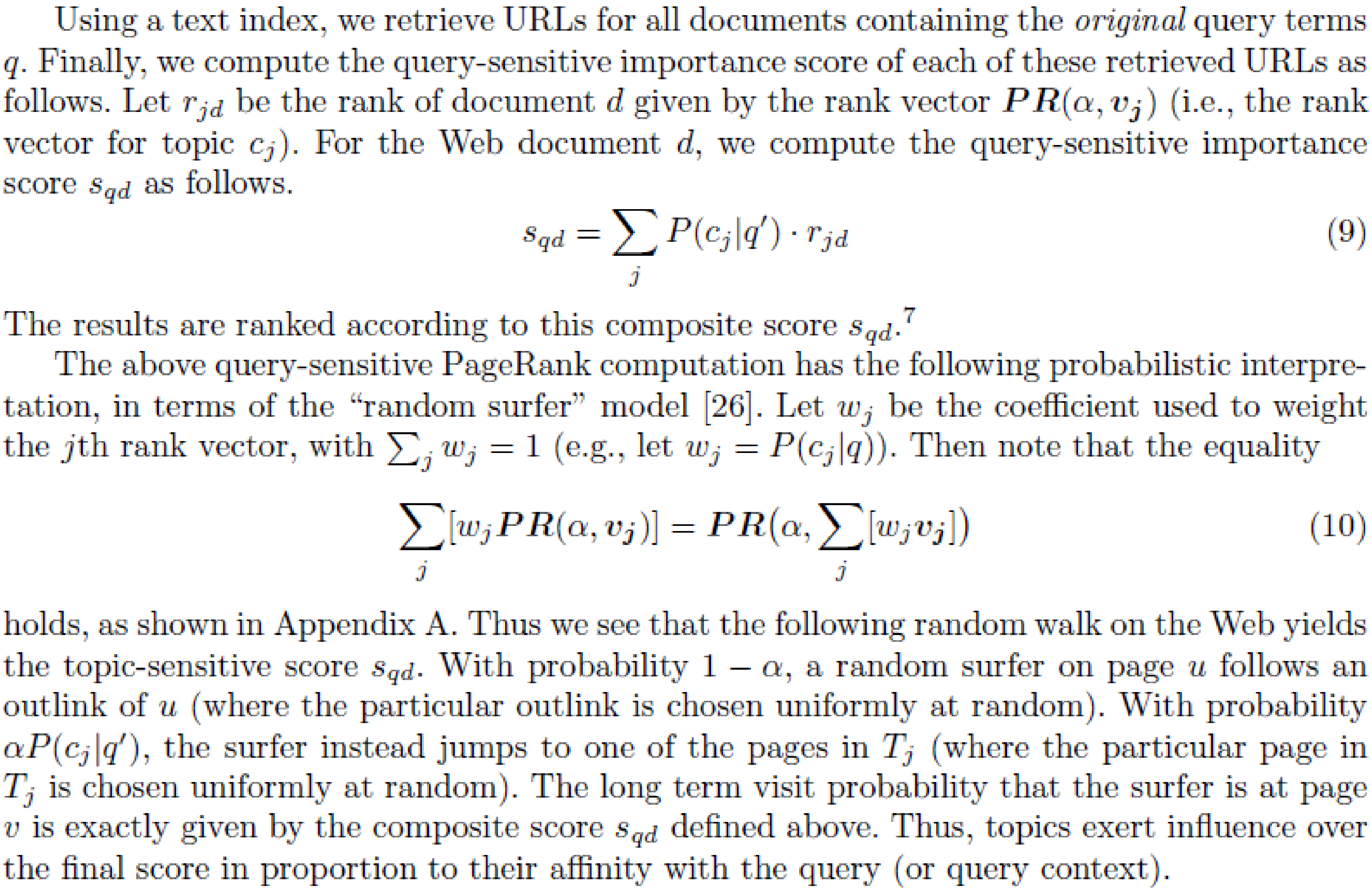
Query Treatment



The second step in our approach is performed at query time. Given a query q, let q' be the context of q. In other words, if the query was issued by highlighting the term q in some Web page u, then q' consists of the terms in u. Alternatively, we could use only those terms in u nearby the highlighted term, as often times a single Web page may discuss a variety of topics. For ordinary queries not done in context, let $q' = q$. Using a multinomial naive-Bayes classifier [24],[6] with parameters set to their maximum-likelihood estimates, we compute the class probabilities for each of the 16 top-level ODP classes, conditioned on q'. Let q'_i be the ith term in the query (or query context) q'. Then given the query q, we compute for each c_j the following:

$$P(c_j|q') = \frac{P(c_j) \cdot P(q'|c_j)}{P(q')} \propto P(c_j) \cdot \prod_i P(q'_i|c_j) \qquad (8)$$

Combining Topics



Using a text index, we retrieve URLs for all documents containing the *original* query terms q. Finally, we compute the query-sensitive importance score of each of these retrieved URLs as follows. Let r_{jd} be the rank of document d given by the rank vector $\boldsymbol{PR}(\alpha, \boldsymbol{v_j})$ (i.e., the rank vector for topic c_j). For the Web document d, we compute the query-sensitive importance score s_{qd} as follows.

$$s_{qd} = \sum_j P(c_j|q') \cdot r_{jd} \tag{9}$$

The results are ranked according to this composite score s_{qd}.[7]

The above query-sensitive PageRank computation has the following probabilistic interpretation, in terms of the "random surfer" model [26]. Let w_j be the coefficient used to weight the jth rank vector, with $\sum_j w_j = 1$ (e.g., let $w_j = P(c_j|q)$). Then note that the equality

$$\sum_j [w_j \boldsymbol{PR}(\alpha, \boldsymbol{v_j})] = \boldsymbol{PR}(\alpha, \sum_j [w_j \boldsymbol{v_j}]) \tag{10}$$

holds, as shown in Appendix A. Thus we see that the following random walk on the Web yields the topic-sensitive score s_{qd}. With probability $1 - \alpha$, a random surfer on page u follows an outlink of u (where the particular outlink is chosen uniformly at random). With probability $\alpha P(c_j|q')$, the surfer instead jumps to one of the pages in T_j (where the particular page in T_j is chosen uniformly at random). The long term visit probability that the surfer is at page v is exactly given by the composite score s_{qd} defined above. Thus, topics exert influence over the final score in proportion to their affinity with the query (or query context).

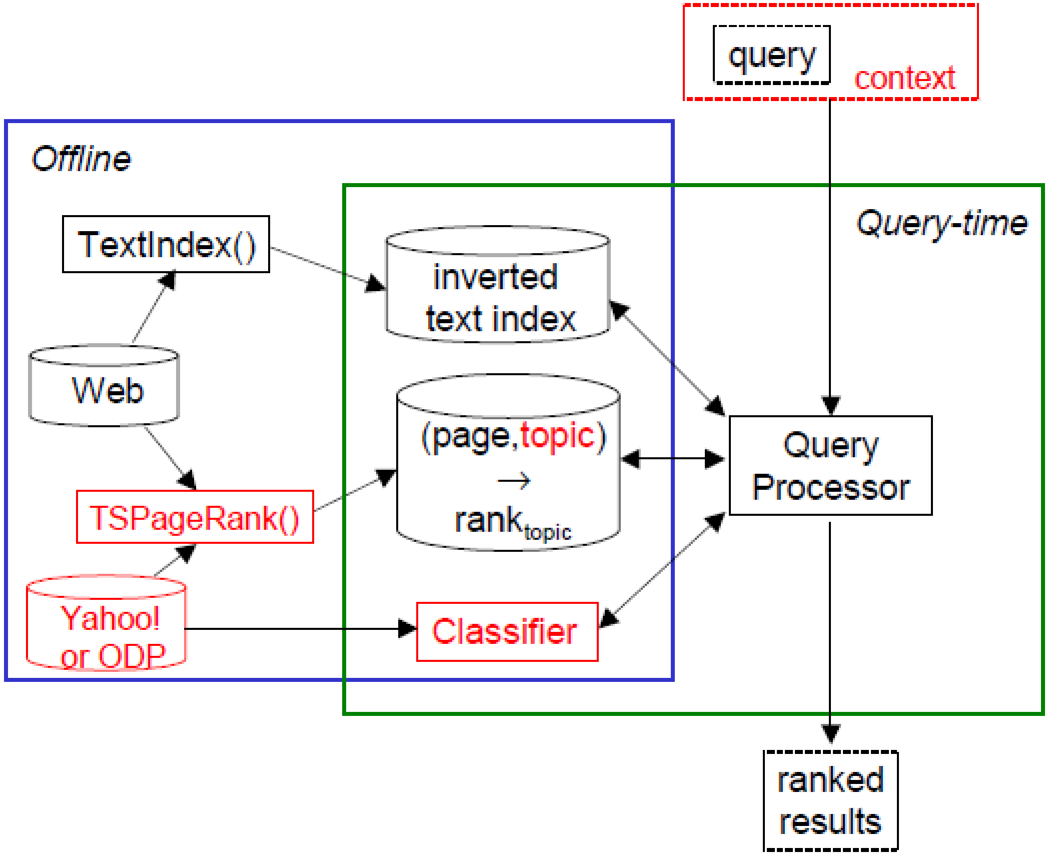


Fig. 2. Illustration of our system utilizing topic-sensitive PageRank.

Table 3. Topic pairs yielding most similar rankings.

Bias-Topic Pair	$OSim$	$KSim$
(Games, Sports)	0.18	0.13
(NoBias, Regional)	0.18	0.12
(Kids & Teens, Society)	0.18	0.11
(Health, Home)	0.17	0.12
(Health, Kids & Teens)	0.17	0.11

Table 5. Estimates for $P(c_j|q)$ for a subset of the test queries.

alcoholism	
Health	0.47
Kids & Teens	0.20
Arts	0.06

bicycling	
Sports	0.52
Regional	0.13
Health	0.07

blues	
Arts	0.52
Shopping	0.12
News	0.08

citrus groves	
Shopping	0.34
Home	0.21
Regional	0.18

classical guitar	
Arts	0.75
Shopping	0.21
News	0.01

computer vision	
Computers	0.24
Business	0.14
Reference	0.09

cruises	
Recreation	0.65
Regional	0.18
Sports	0.04

death valley	
Regional	0.28
Society	0.14
News	0.10

field hockey	
Sports	0.89
Shopping	0.03
Reference	0.03

graphic design	
Computers	0.36
Business	0.23
Shopping	0.09

gulf war	
Society	0.21
Kids & Teens	0.18
Regional	0.17

hiv	
Health	0.40
News	0.19
Kids & Teens	0.14

java	
Computers	0.53
Games	0.10
Kids & Teens	0.06

lyme disease	
Health	0.96
Regional	0.01
Recreation	0.01

mutual funds	
Business	0.77
Regional	0.05
Home	0.05

parallel architecture	
Computers	0.70
Science	0.10
Reference	0.07

rock climbing	
Recreation	0.54
Regional	0.13
Sports	0.07

san francisco	
Sports	0.27
Regional	0.16
Recreation	0.10

shakespeare	
Arts	0.34
Reference	0.21
Kids & Teens	0.15

table tennis	
Sports	0.53
Shopping	0.14
Regional	0.09

telecommuting	
Business	0.70
Kids & Teens	0.04
Society	0.03

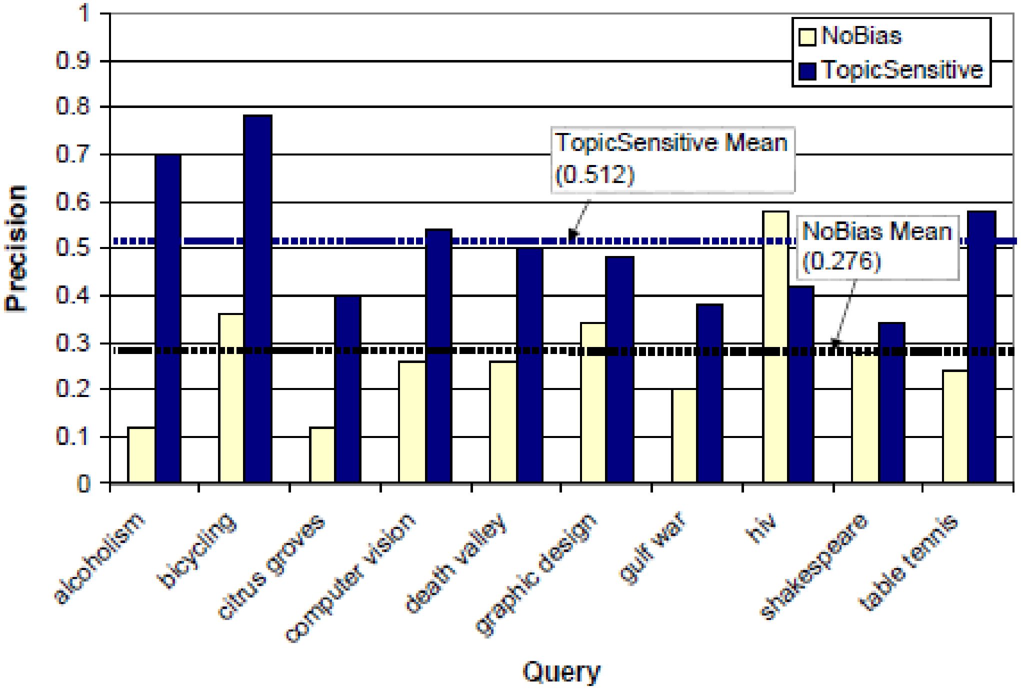


Fig. 3. Precision @ 10 results for our test queries. The average precision over the ten queries is also shown.

Applications

- Domain specific PageRank
- User Profiling, when **p** is based on user preferred topics
- Semantic Classification:
  - Word Sense Disambiguation (Soroa & Agirre, 2009) when:
    - The graph is a sense dictionary (i.e. Wordnet)
    - **PR** gives the preference to a sense as its reachability
    - **p** is based on the context of the target word

References

- Taher H. Haveliwala, *Topic-Sensitive PageRank* In Proceedings of the Eleventh International World Wide Web Conference, May 2002.

- E. Agirre and A. Soroa. 2009. *Personalizing pagerank for word sense disambiguation*. in Proceedings of EACL-09, Athens, Greece.