Presentation of

“Topic-Sensitive PageRank”

Taher H. Haveliwala

Lucas Panjer

November 23, 2006

Concept

- PageRank provides a general “importance” of a web page without the context of a query
- Compute a set of PageRank bias vectors by topic
- Gather topics from Open Directory

Usage

- User submits a query
- Topic is selected based on context:
  - Query terms
  - Topic of current page
  - Past query history
- Selected topics are weighted more heavily

Approach

- Precomputation
  - For each of the 16 top-level ODP categories
    - For each page
      - Generate a PageRank vector
      - Generate a class term vector
  - ODP chosen as it is relatively free from classification bias and edited by many contributors

Approach (2)

- Query-time
  - – Determine a context
      - » Highlighted term
      - » Search term
    - Compute the probability of the topic given a context
  - – Query sensitive PageRank score is probability of a topic given a context x PR vector for the topic

$$s_{qd} = \sum_{j} P(c_j|q') \cdot rank_{jd}$$

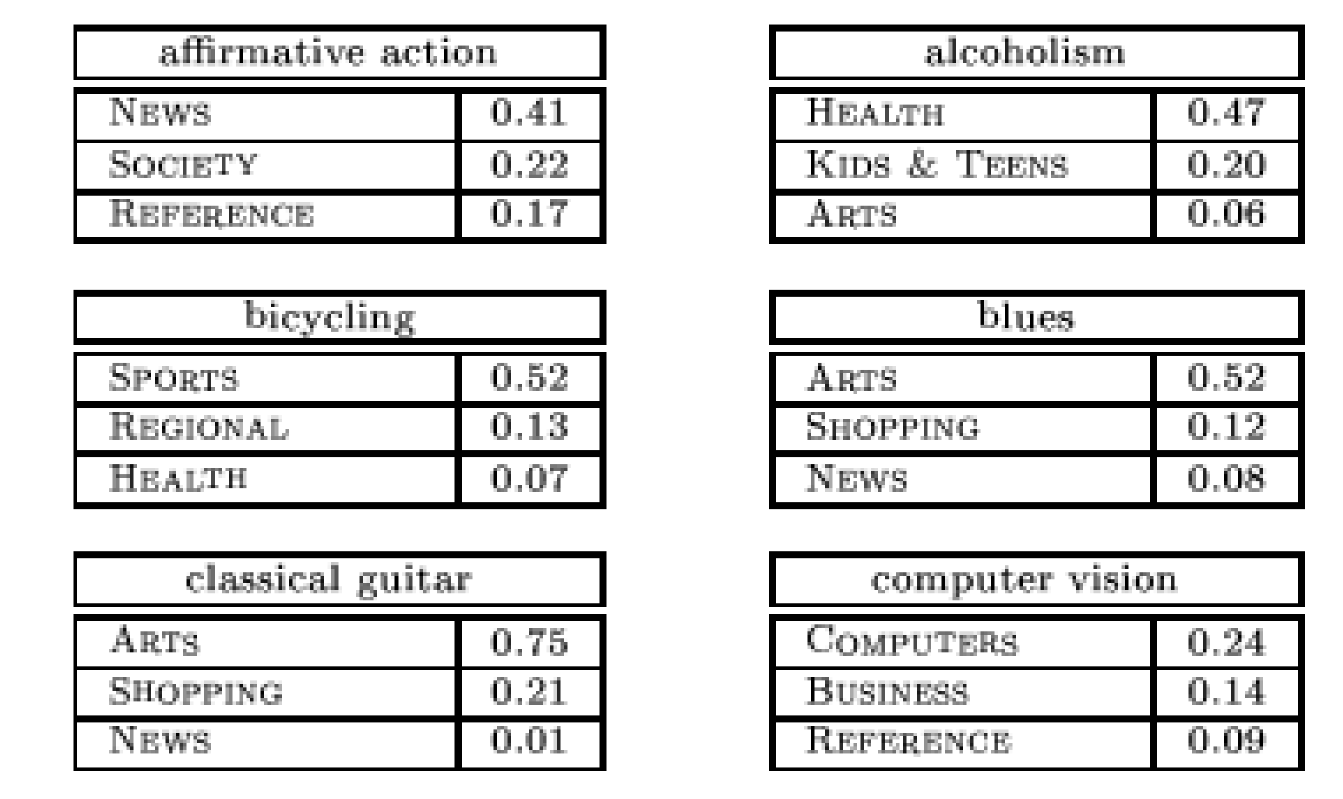
Experimentation

- 35 test queries from previous paper
- Dataset from Stanford WebBase contained 280k of 3M available ODP listed pages
- Queried each of the 16 topics indicies and a NOBIAS index
- Similarity of result ordering
  - OSim : degree of overlap between top 20 URLs
  - KSim : Kendall’s distance, number of pairwise swaps necessary to align two lists

Experimentation (2)

- User study
  - 5 volunteers
  - 10 queries randomly selected from set of 35
  - Volunteer shown NOBIAS and biased ranking
    - Selected all URLs in result set which were “relevant” to the query
    - Selected the better ranking

Results



affirmative action	
News	0.41
Society	0.22
Reference	0.17

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

classical guitar	
Arts	0.75
Shopping	0.21
News	0.01

computer vision	
Computers	0.24
Business	0.14
Reference	0.09

- Probability of a topic given a query

Results (2)

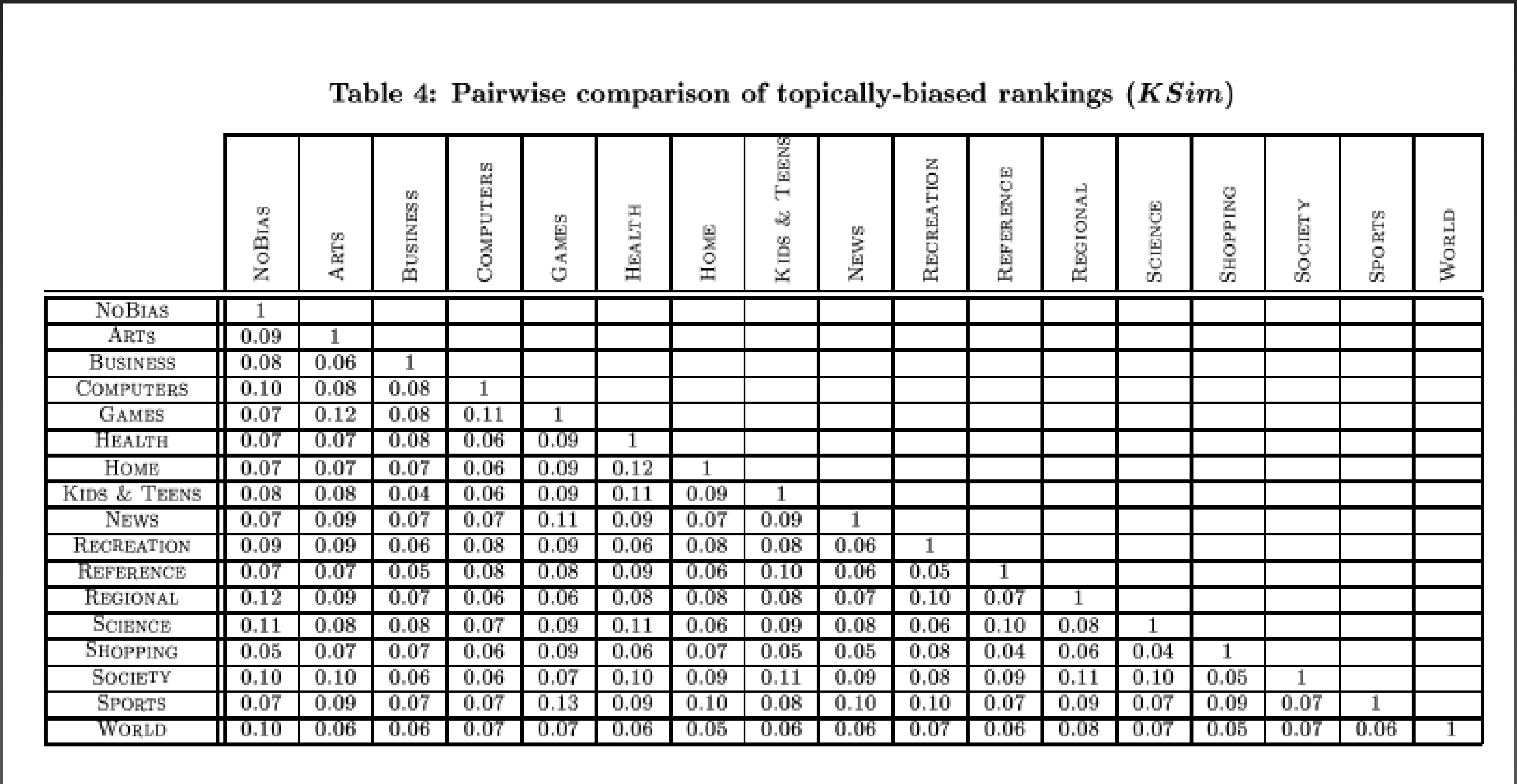


Table 4: Pairwise comparison of topically-biased rankings (***KSim***)

	NoBias	Arts	Business	Computers	Games	Health	Home	Kids & Teens	News	Recreation	Reference	Regional	Science	Shopping	Society	Sports	World
NoBias	1																
Arts	0.09	1															
Business	0.08	0.06	1														
Computers	0.10	0.08	0.08	1													
Games	0.07	0.12	0.08	0.11	1												
Health	0.07	0.07	0.08	0.06	0.09	1											
Home	0.07	0.07	0.07	0.06	0.09	0.12	1										
Kids & Teens	0.08	0.08	0.04	0.06	0.09	0.11	0.09	1									
News	0.07	0.09	0.07	0.07	0.11	0.09	0.07	0.09	1								
Recreation	0.09	0.09	0.06	0.08	0.09	0.06	0.08	0.08	0.06	1							
Reference	0.07	0.07	0.05	0.08	0.08	0.09	0.06	0.10	0.06	0.05	1						
Regional	0.12	0.09	0.07	0.06	0.06	0.08	0.08	0.08	0.07	0.10	0.07	1					
Science	0.11	0.08	0.08	0.07	0.09	0.11	0.06	0.09	0.08	0.06	0.10	0.08	1				
Shopping	0.05	0.07	0.07	0.06	0.09	0.06	0.07	0.05	0.05	0.08	0.04	0.06	0.04	1			
Society	0.10	0.10	0.06	0.06	0.07	0.10	0.09	0.11	0.09	0.08	0.09	0.11	0.10	0.05	1		
Sports	0.07	0.09	0.07	0.07	0.13	0.09	0.10	0.08	0.10	0.10	0.07	0.09	0.07	0.09	0.07	1	
World	0.10	0.06	0.06	0.07	0.07	0.06	0.05	0.06	0.06	0.07	0.06	0.08	0.07	0.05	0.07	0.06	1

- All topics are substantially different

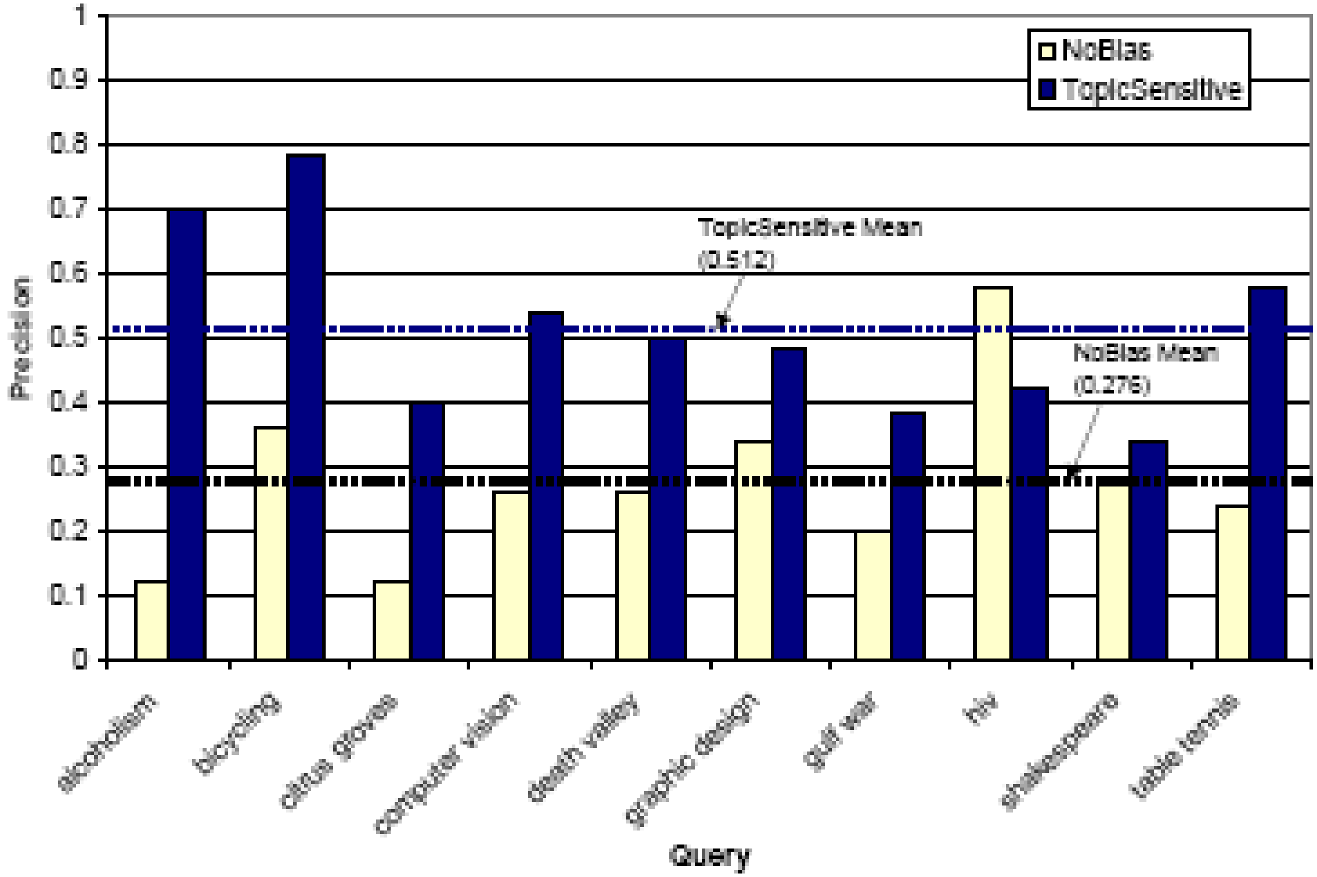
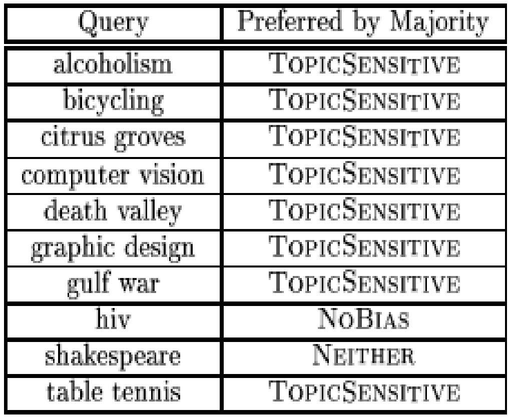


Figure 1: Precision @ 10 results for our test q
The average precision over the ten queries
shown.

- Precision : Fraction of top 10 URLs deemed relevant by user

Table 7: Ranking preferred by majority of users



Query	Preferred by Majority
alcoholism	TopicSensitive
bicycling	TopicSensitive
citrus groves	TopicSensitive
computer vision	TopicSensitive
death valley	TopicSensitive
graphic design	TopicSensitive
gulf war	TopicSensitive
hiv	NoBias
shakespeare	Neither
table tennis	TopicSensitive

Contributions

- Refining PageRank
- Combining categorization and search
    - directory + index
- Quick customization of search results to context

Positive

- Simple to customize a search engine for a specific collection, can make contexts from any URL set
- Simple one time calculations
- (Significantly) better results

Negative

- User study could be much more broad
- Outside of a well controlled context URL set this could be manipulated perhaps even more than PageRank
- Most common topic for a query context can overshadow other topics