



Topic-Sensitive PageRank

Taher H. Haveliwala
Stanford University

taherh@cs.stanford.edu

Motivation

- Improve search results
 - Current engines work well for us “computer types”, but not for novice users
- Exploit search context in a tractable and effective way
 - Current engines can only do so well when optimizing parameters for Joe User issuing query q

2

Search Context

- Query context
 - Highlighted word on page
 - Previous queries issued
- User context
 - Bookmarks
 - Browsing history
- *Placing Search in Context: The Concept Revisited*
 - [Finkelstein et al. WWW10 '01]

3

Link-Based Scoring (HITS)

- HITS (“Hubs and Authorities”)
 - [Kleinberg SODA '98]
 - Determine important *Hub* pages and important *Authority* pages
 - +Query specific rank score
 - - Expensive at runtime

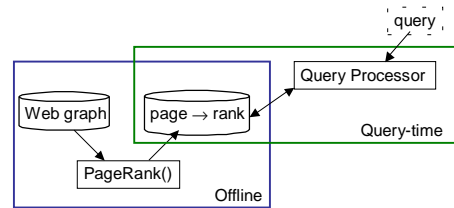
4

Link-Based Scoring (PageRank)

- PageRank
 - [Page et al. '98]
 - Assigns a-priori "importance" estimates to pages
 - - Query independent rank score
 - + Inexpensive at runtime
- Algorithm has hooks for "personalization"

5

Original PageRank



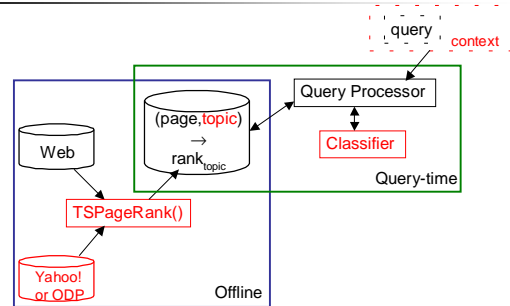
6

Topic-Sensitive PageRank

- Assigns *multiple* a-priori "importance" estimates to pages
- One PageRank score per *basis topic*
 - + Query specific rank score
 - + Make use of context
 - + Inexpensive at runtime
- Related approach: one score per query word was considered in [Richardson, Domingos NIPS '02] (builds on [Rafiei, Mendelzon WWW '00])

7

Topic-Sensitive PageRank



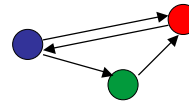
8

Original PageRank Intuition

- *“Page is important if many important pages point to it”*
 - Many pages point to Yahoo!, so it is “important”
 - Because Yahoo! is important, anyone it prominently points to is “important”

9

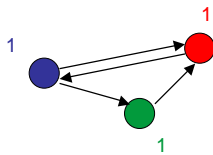
PageRank Diagram



Graph structure for *entire* web

11

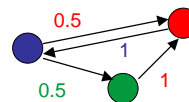
PageRank Diagram



Initialize all nodes to rank 1

12

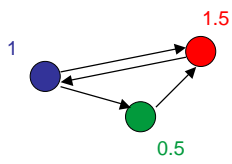
PageRank Diagram



Propagate ranks across links
(multiplying by link weights)

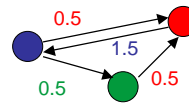
13

PageRank Diagram



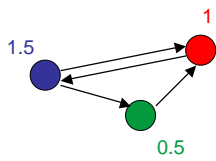
14

PageRank Diagram



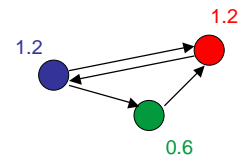
15

PageRank Diagram



16

PageRank Diagram



After a while...

17

Original PageRank

- Input
 - Web graph G
- Output
 - Rank vector r : (page \rightarrow page importance)
- $r = PR(G)$

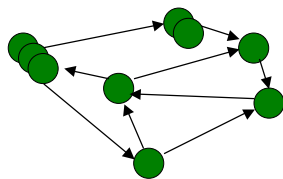
18

Influencing the Computation

- Uninfluenced:
"Page is important if many important pages point to it."
- Influenced:
"Page is important if many important pages point to it, and btw, the following are by definition important pages."

19

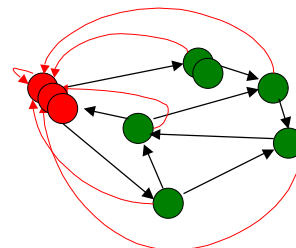
Influencing the Computation



Graph structure for *entire* web

20

Influencing the Computation



Pick a **set** of influence

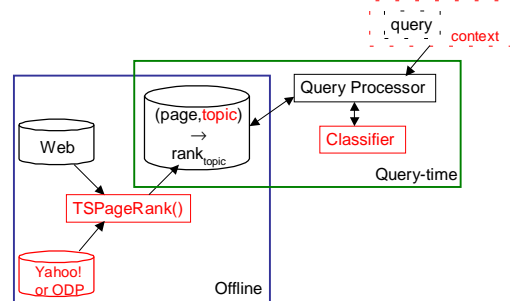
21

Influenced PageRank

- Input:
 - Web graph G
 - influence vector \mathbf{v}
 $\mathbf{v} : (\text{page} \rightarrow \text{degree of influence})$
- Output:
 - Rank vector $\mathbf{r} : (\text{page} \rightarrow \text{page importance wrt } \mathbf{v})$
- $\mathbf{r} = \text{IPR}(G, \mathbf{v})$
- How to choose \mathbf{v} ?

22

Topic-Sensitive PageRank



23

Topic-Sensitive PageRank: Part I (preprocessing)



- Goal: Generate *multiple* a-priori estimates of page importance, each score providing an importance estimate with respect to a *topic*
- Use the Open Directory as a source of representative *basis* topics (i.e., use ODP pages to form a set of influence vectors \mathbf{v}_j)
- Offline preprocessing step, just as with ordinary PageRank

24

Offline Processing

- Input:
 - Web W
 - Basis topics $[c_1, \dots, c_{16}]$
 We use 16 categories (first level of ODP)
- Output:
 - List of rank vectors $[\mathbf{r}_1, \dots, \mathbf{r}_{16}]$
 $\mathbf{r}_j : (\text{page} \rightarrow \text{page importance wrt topic } c_j)$

25

Offline Processing

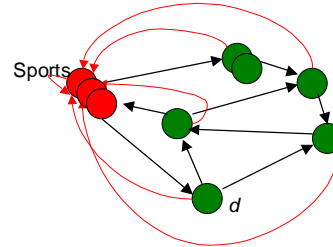
For each topic $c_j \in \text{FirstLevel(ODP)}$:

$$\text{set } v_j[i] = \begin{cases} \frac{1}{|\text{pages}(c_j)|} & \text{if } i \in \text{pages}(c_j) \\ 0 & \text{otherwise} \end{cases}$$

Compute $r_j = \text{IPR}(W, v_j)$

26

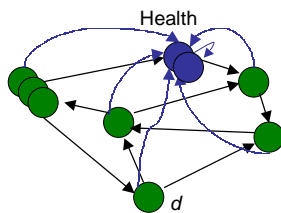
Graphical Depiction of Part I



Select set of influence, calculate PageRank for all pages
For example, $r_{sports}[d] = .05$

27

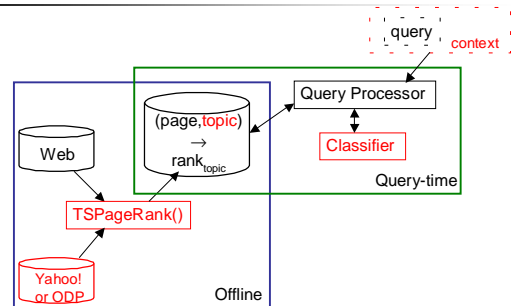
Graphical Depiction of Part I



Select set of influence, calculate PageRank for all pages
For example, $r_{health}[d] = .01$

28

Topic-Sensitive PageRank



29

Topic-Sensitive PageRank: Part II (query processing)



- Goal: calculate some distribution of weights over the 16 topics in our basis
- Use a multinomial Naive Bayes classifier
 - Training set: pages listed in ODP
 - Input: {query} or {query, context}
 - Output: probability distribution (weights) over the basis topics

30

Two Usage Scenarios

- Classify the query
- Classify the query + context
 - query history
 - words surrounding a highlighted search phrase
 - ...

31

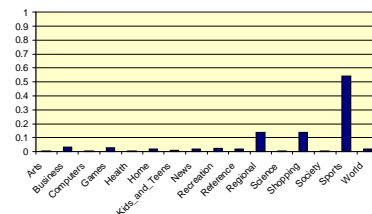
Classify the Query

- Only the link structure of pages relevant to the query topic will be used to rank pages
- Better to rank query 'golf' with the Sports-specific rank vector

32

Example Topic Distribution

- For the query 'golf', with no additional context, the distribution of topic weights we would use is:



33

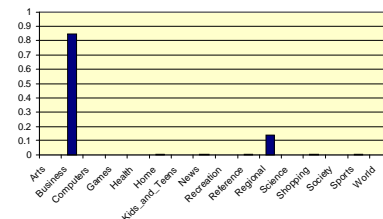
Classify the Query Context

- The topic distribution will influence rankings to prefer pages important to the topic of the query context
- If user issues queries about investment opportunities, a follow-up query on 'golf' should be ranked with the Business-specific rank vector

34

Picking the Topic Distribution

- If the query is 'golf', but the previous query was 'financial services investments', then the distribution of topic weights we would use is:



35

Composite Link Score

- Use the distribution \mathbf{w} to weight the respective topic-specific ranks, forming the topic-sensitive PageRank score for document d :

$$s_d = \sum_j w_j r_j[d]$$

36

Interpretation of Composite Score

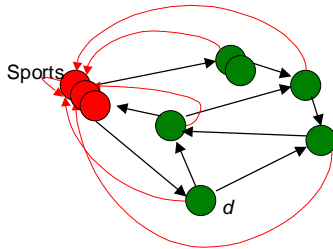
- For set of influence vectors $\{\mathbf{v}_j\}$

$$\sum_j [w_j \cdot \text{IPR}(W, \mathbf{v}_j)] = \text{IPR}(W, \sum_j [w_j \cdot \mathbf{v}_j])$$

- Weighted sum of rank vectors itself forms a valid rank vector

37

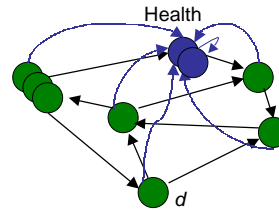
Interpretation



First set of influence

38

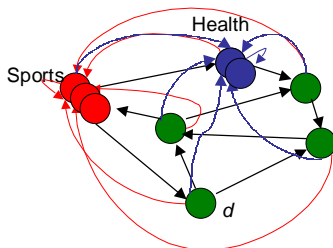
Interpretation



Second set of influence

39

Interpretation



Topic-sensitive score is PageRank of above graph
For example, $r_{\{sports, health\}, d} = .026$

40

Implementation Platform

- Stanford WebBase repository: 120M pages
- For research experiments, topic weights can be estimated automatically by classifier, or specified explicitly

41

Does it make a difference?

- Do the different topical rank vectors rank results for queries differently?
- To answer, measure the similarity of induced ranks for some set of test query results
- Details in paper, but short answer is, "yes, the different rank vectors induce different result rankings"

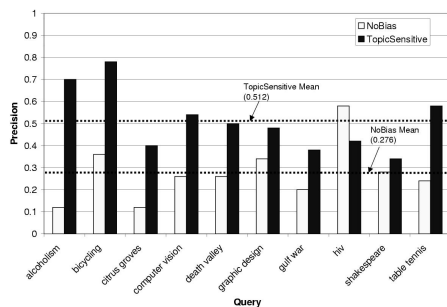
42

User Study (no search context)

- Test set of 10 queries
- 5 users were each shown top 10 results to queries, when ranked using
 - Standard PageRank vector
 - Topic-Sensitive PageRank vector
- A page in the result was "relevant" if 3 of the 5 users judged it to be relevant

43

User Study (no search context)



44

User Study Follow-up

- After factoring in text-based scoring, the precision values for both standard and topic-sensitive ranking go up
- Topic-sensitive rankings still preferred
- "Precision" not the best metric to use
 - Some pages are "more relevant"
 - Some pages are of "higher quality"



45

Query for 'golf' (topic-sensitivity disabled)

The screenshot shows a web browser window titled "Index demo - Microsoft Internet Explorer". The address bar is empty. The main content area has the heading "Index demo" and a search form with the text "Enter search query:" and a search box containing the word "golf". Below the search box, there are four checkboxes under the heading "Select sources of context below:":
 Enable auto detect of topic
 Query history
 Bookmarks
 Browsing history

46

Results for 'golf'

The screenshot shows a web browser window titled "Web Demo - Results for 'golf' - Microsoft Internet Explorer". The main content area has the heading "TSPR Demo: +golf" and a list of search results:

- 13 7208 www.golf.com
<http://www.golf.com/> cache
- 12 8922 [LI GOLF COM - "For Sale" Postings TOC](http://www.lgolf.com/discuss/for-sale)
http://www.lgolf.com/discuss/for-sale_toc.htm cache
- 12 8168 [World Golf Village, Home of the World Golf Hall of Fame, St. Augustine, FL](http://www.wgv.com/)
<http://www.wgv.com/> cache
- 12 1651 [GOLFONLINE, FROM THE EDITORS OF GOLF MAGAZINE](http://www.golfonline.com/)
<http://www.golfonline.com/> cache

47

Enable History Tracking 'financial services investments'

The screenshot shows a web browser window titled "Index demo - Microsoft Internet Explorer". The address bar is empty. The main content area has the heading "Index demo" and a search form with the text "Enter search query:" and a search box containing the text "financial services investments". Below the search box, there are four checkboxes under the heading "Select sources of context below:":
 Enable auto detect of topic
 Query history
 Bookmarks
 Browsing history

48

Results 'financial services investments'

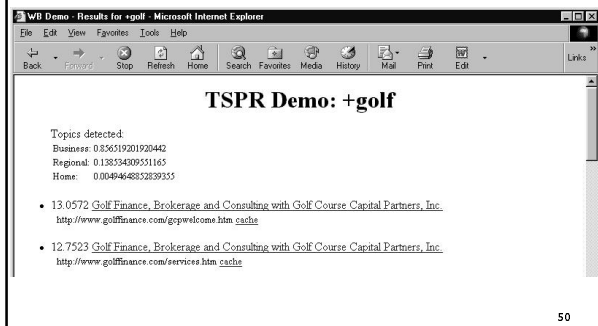
The screenshot shows a web browser window titled "Web Demo - Results for 'financial services investments' - Microsoft Internet Explorer". The main content area has the heading "TSPR Demo: +financial +services +investments" and the following text:

Topics detected:
Business: 0.953822154259855
Regional: 0.63646731590953028
Reference: 0.00971052673483896

- 15 151 [FCNB Bank, Frederick Maryland, Online Banking, Loans, CD, Mortgages, Business Services, Investments, Trusts, Insurance, Financial Calculators, Checking, Home banking, Web Banking](http://www.fcnb.com/)
<http://www.fcnb.com/index.htm> cache
- 14 4445 [Equity Group Services](http://www.theequitygroup.com/serv.html)
<http://www.theequitygroup.com/serv.html> cache
- 14 3376 [Allegheym Financial Group, Ltd](http://www.allegheymfinancial.com/pages/investment.html)
<http://www.allegheymfinancial.com/pages/investment.html> cache

49

'golf' again, but query history judged to be *Business* topic



The screenshot shows a Microsoft Internet Explorer browser window with the title "WB Demo - Results for 'golf' - Microsoft Internet Explorer". The address bar shows "http://www.golffinance.com/gp/welcome.htm". The main content area displays "TSPR Demo: +golf" and "Topics detected:" followed by a list of topics and their associated URLs. The topics are:

- Business: 0.836519201920442
- Regional: 0.136354809301165
- Home: 0.00494648832839355

Below the topics, there are two search results:

- 13.0572 Golf Finance, Brokerage and Consulting with Golf Course Capital Partners, Inc. <http://www.golffinance.com/gp/welcome.htm> cache
- 12.7523 Golf Finance, Brokerage and Consulting with Golf Course Capital Partners, Inc. <http://www.golffinance.com/services.htm> cache

The page number "50" is visible in the bottom right corner.

Search Context

- Advantages of mediating through basis topics, as opposed to 'keyword extraction':
 - **Flexibility:** uniformly treat variety of sources of context and personalization
 - **Transparency:** topic weights are easily interpreted by user
 - **Privacy:** topic weights reveal less unintentionally
 - **Efficiency:** low query time cost, with small additional preprocessing cost

51

Future Work

- Finer grained set of representative topics, to reflect more accurately user preferences and search context



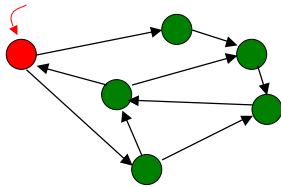
52

Future Work

- Graph weighting scheme based on page *similarity* to ODP category, rather than page *membership* to ODP category

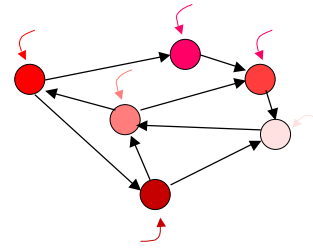
53

Current Approach



54

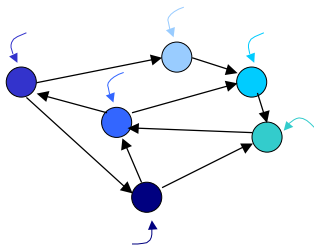
Alternative Approach



Sports

55

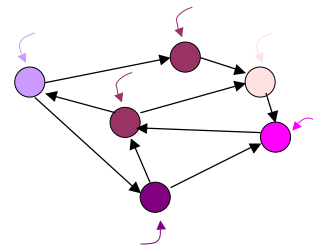
Alternative Approach



Health

56

Alternative Approach



{Sports,Health}

57

Related Work

- *Scaling Personalized Search*
 - [Jeh,Widom '02]
 - Dynamic programming for generation of complete basis
- *What is this Page Known For?*
 - [Rafiei,Mendelzon WWW9 '00]
 - What keywords is a page known for?
- *The Intelligent Surfer: ...*
 - [Richardson,Domingos NIPS '02]
 - Computes PageRank once for each query
- *Persona*
 - [Tanudjaja,Mui HICSS '02]
 - Enhances HITS with ODP data

58