Evaluating Strategies for Similarity Search on the Web

Taher H. Haveliwala
Aristides Gionis
Dan Klein
Piotr Indyk
{taherh,gionis,klein}@cs.stanford.edu
indyk@theory.lcs.mit.edu

Similarity search

- Given a query Web page *q*, return Web pages that are "similar" to *q*

www.moneycentral.com
www.pathfinder.com/money
www.moneyworld.co.uk
www.money.com
www.etrade.com
www.moneyclub.com

2

Similarity search

- Two major issues:
  - Choose the strategy that best captures the notion of Web-page "similarity"
  - Scaling up the chosen strategy to repository of millions of pages

3

Related work

- *Finding Related Pages in the WWW*
  - [Dean,Henzinger WWW8 '99]
- *Automatic Resource Compilation ...*
  - [Chakrabarti et al WWW7 '98]
- Commercial search engines

4

Model for document similarity

- Represent each Web page as bag of terms
  - content, anchor-text, links, ...
- Similarity of two pages is given by similarity their respective bags
  - cosine
  - **Jaccard**

5

Model for document similarity

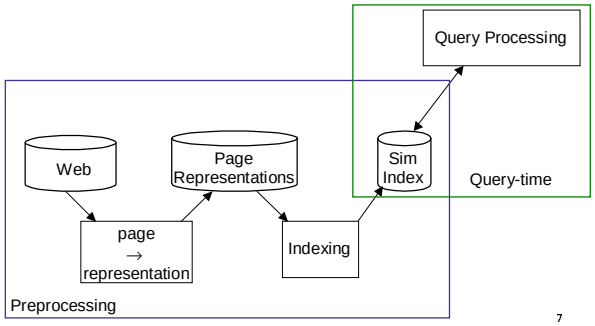
- For pages *a* and *b*, with respective bags α and β, define

$$sim(a,b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|}$$

- Strategy for (page $\rightarrow$ bag) is the crucial step in quality of *sim()*

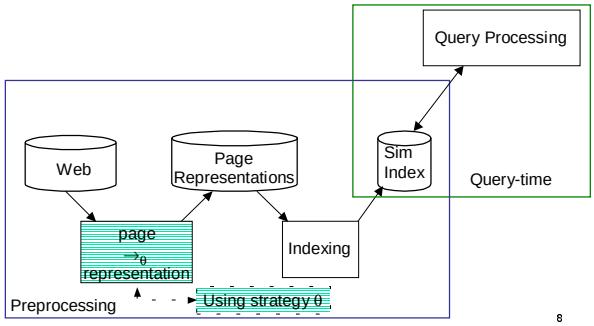
6

Similarity search system

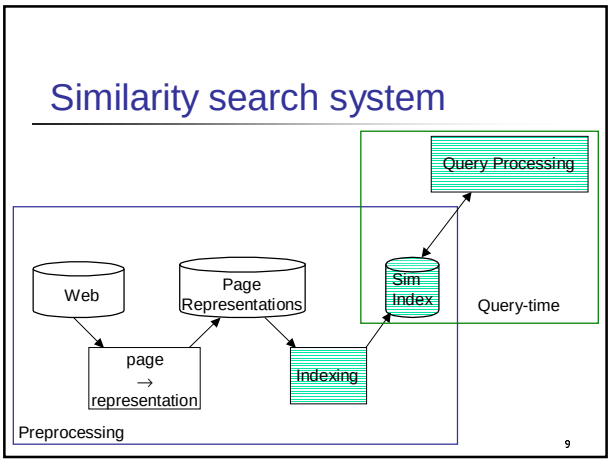


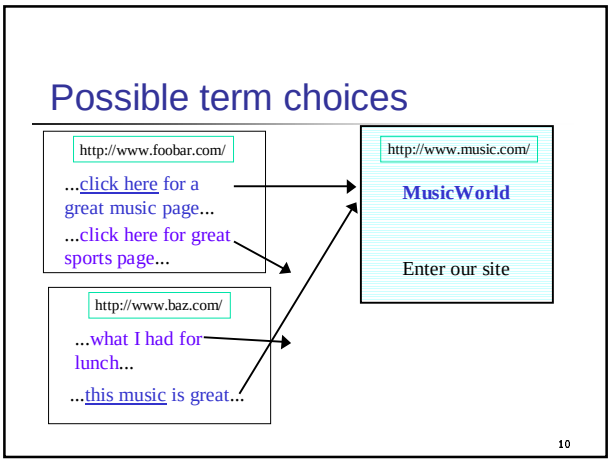
7

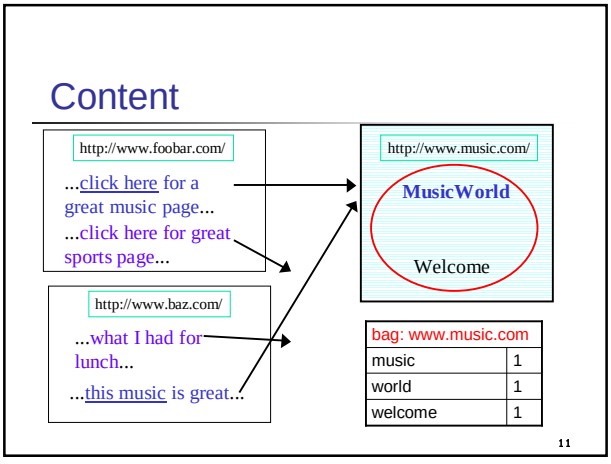
Similarity search system

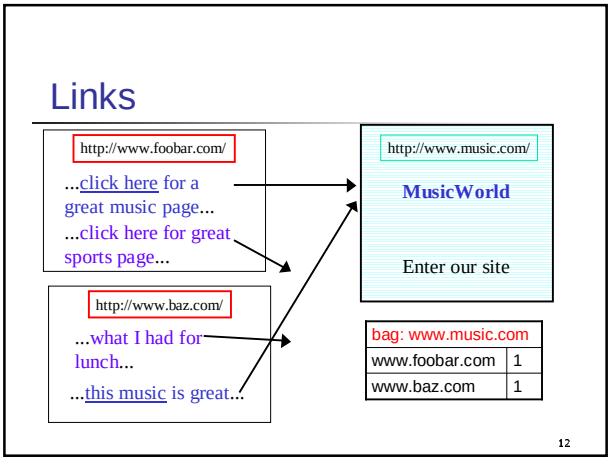


8

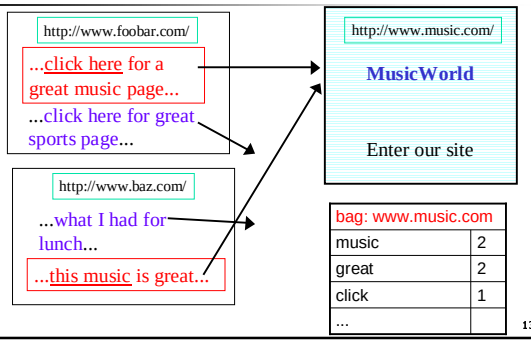

Similarity search system
Query Processing
Web
Page Representations
Sim Index
Query-time
page → representation
Indexing
Preprocessing
9


Possible term choices
http://www.foobar.com/
...click here for a great music page...
...click here for great sports page...
http://www.baz.com/
...what I had for lunch...
...this music is great...
http://www.music.com/
MusicWorld
Enter our site
10


Content
http://www.foobar.com/
...click here for a great music page...
...click here for great sports page...
http://www.baz.com/
...what I had for lunch...
...this music is great...
http://www.music.com/
MusicWorld
Welcome
bag: www.music.com
music 1
world 1
welcome 1
11


Links
http://www.foobar.com/
...click here for a great music page...
...click here for great sports page...
http://www.baz.com/
...what I had for lunch...
...this music is great...
http://www.music.com/
MusicWorld
Enter our site
bag: www.music.com
www.foobar.com 1
www.baz.com 1
12

Anchor windows



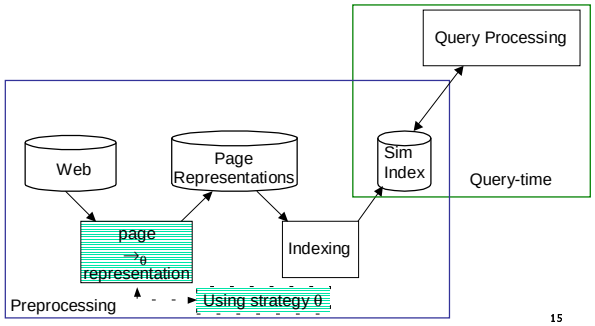
13

Parameter space for bag generation

- Space of parameters considered:
  - content vs. links vs. anchor windows
  - anchor window length
  - term weighting schemes
- Choice of a particular assignment of parameters, θ, defines a similarity search strategy

14

Similarity search system



15

(Strategy, query) → similarity ordering

- Inputs:
  - $\theta \in \Theta$: strategy (i.e., parameter setting)
  - $q \in$ Web: query page
- Outputs:
  - τ: list of web pages ordered by similarity to q using strategy θ
- $\tau = T(\theta, q)$

16

Evaluating strategies

- Goal: find "best" $\theta_i \in \Theta$
- Develop system to measure quality of different parameter settings
  - What do you choose as the ground truth for Web-page similarity?
  - How do you compare a particular strategy to this ground truth?

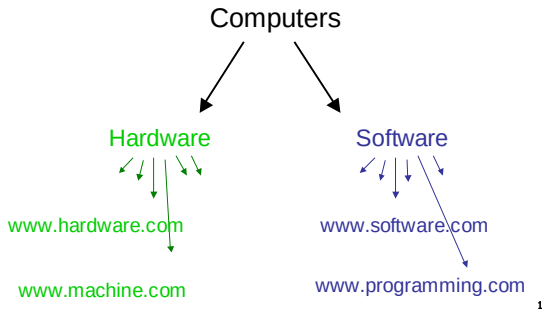
17

Web directories (Yahoo!, ODP)

- Hand-constructed hierarchical directories such as Yahoo! and the Open Directory Project (ODP) can be used as an external quality measure
- Do not directly provide ranked similarity listings
- Do contain many implicit similarity judgements

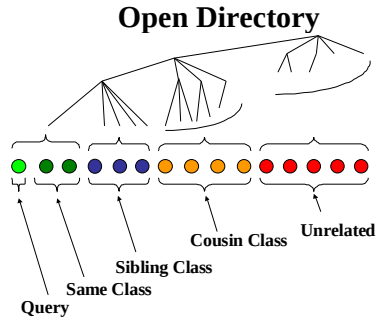
18

Directory → Similarity judgements



19

(Directory, query) → similarity ordering



20

(Directory, query) → similarity ordering

- Inputs:
  - **D**: hierarchical directory
  - $q \in$ **D**: query page
- Outputs:
  - τ: list of pages of **D** partially ordered by similarity to *q*, using the ordering implicit in **D**
- $\tau = T(\mathbf{D}, q)$
- The above is for *evaluating* similarity search, not performing it!

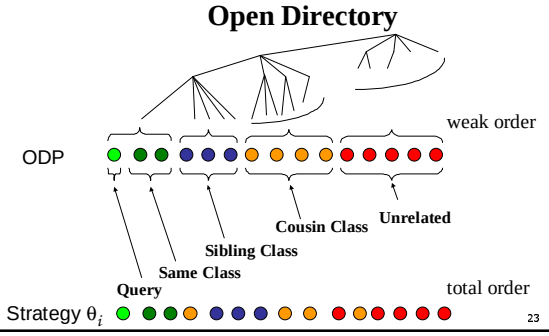
21

Evaluating strategies

1. Restrict attention during evaluation phase to pages in the directory **D**
2. Compare similarity ordering induced by parameter setting θ_i to the similarity ordering induced by the directory, over test set of query pages
3. Choose the θ_i that agrees most closely with the judgements in **D**

22

Directory vs. Strategy



23

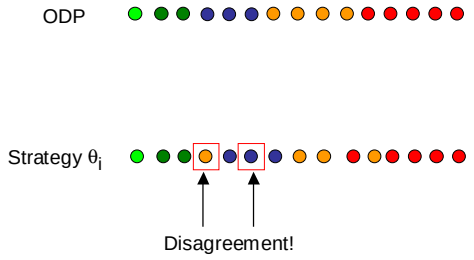
Comparing two orderings

- Based on Kruskal-Goodman Γ
- Inputs
  - τ_{odp}: strict weak ordering of pages (ODP)
  - τ_i : total ordering of pages according to θ_i
- Output
  - $-1 \le \Gamma \le 1$: measure of agreement

    $2 \times \Pr[\tau_{odp} \text{ and } \tau_i \text{ agree on ordering of } (u,v)] - 1$

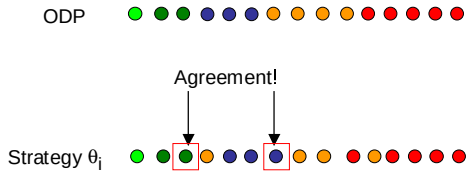
24

Directory vs. Strategy



25

Directory vs. Strategy



26

Example of two rankings with different Γ scores

Query page: www.aabga.org
(American Association of Botanical Gardens and Arboreta)

Canadian Botanical Conservation Network http://www.rbg.ca/cbcn	The Huntington Library, Art Collections, and Botanical Gardens www.huntington.org
The Royal Horticultural Society www.rhs.org.uk	The American Rhododendron Society www.rhododendron.org
The American Rhododendron Society http://www.rhododendron.org	American Chiropractic Association www.amerchiro.org
Gardener's Supply Company www.vg.com	American Trakehner Association (horses) www.americantrakehner.com
The New England Botanical Club www.herbaria.harvard.edu/collections/nebc/nebc.html	American Subcontractors Association www.asaonline.com

Γ=0.5312 Γ=0.3096

27

Evaluating strategies

1. For each $\theta_i \in \Theta$
   $\Gamma_{\theta_i} = Avg_{q \in \mathbf{D}} [\Gamma(T(\mathbf{D}, q), T(\theta_i, q))]$
2. Select strategy $\theta^* = argmax_{\theta_i} [\Gamma_{\theta_i}]$

Only assumes that higher agreement, *on average,* with ODP is a good thing

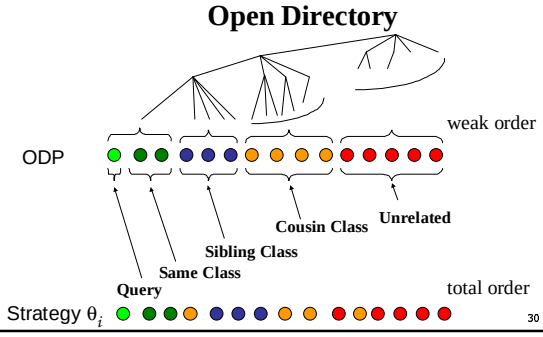
28

Experimental results

- 42 million page subset of the Web from the Stanford WebBase
- Following results restrict attention to two colors: same class and sibling class
- D: 300 pairs of sibling clusters from ODP

29

Directory vs. Strategy



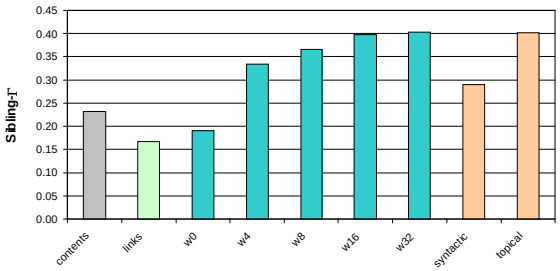
30

Feature space: term selection

- Content
- Inlinks
- Anchor-windows
  - Basic
    - window size W ∈ {0,4,8,16,32}
  - Syntactic
    - averaged 3 words in both directions
  - Topical
    - averaged 21 words in both directions

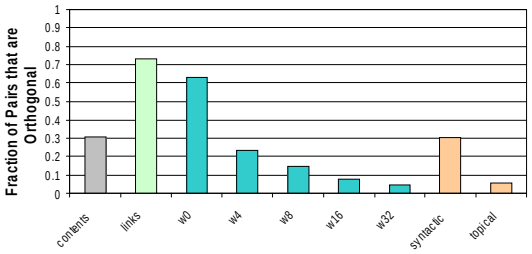
31

Γ scores



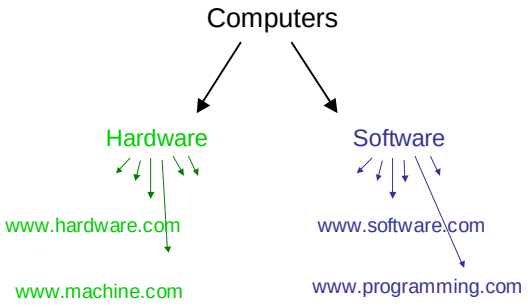
32

Orthogonality



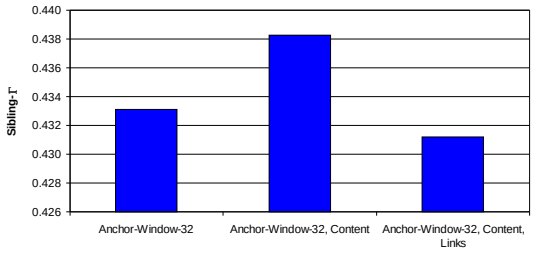
33

Directory → Similarity judgements



34

Composite schemes

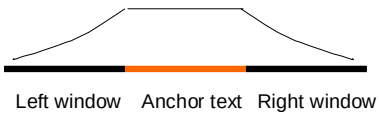


Note: distance weighting was enabled

35

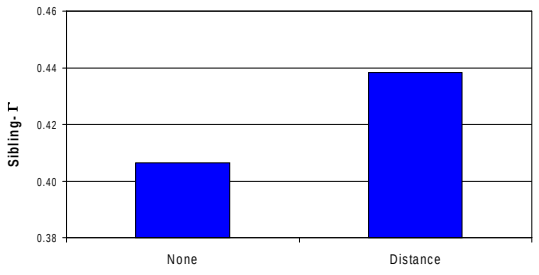
Feature space: term weighting

- Distance weighting for anchor-window terms



36

Weighting schemes



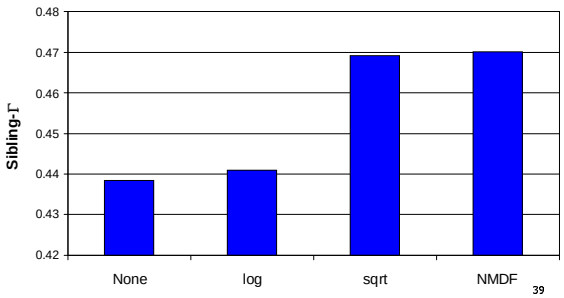
37

Feature space: term weighting

- Frequency based weighting schemes
  - Inverse Document Frequency (IDF)
    - attenuate weights for frequent terms
  - Nonmonotonic Document Frequency (NMDF)
    - attenuate weights for frequent and infrequent terms

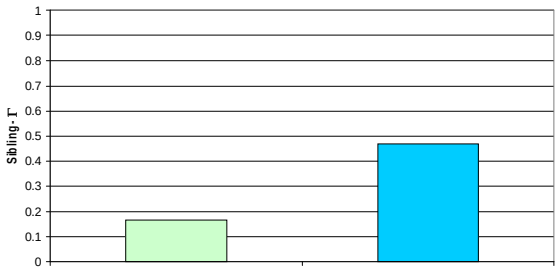
38

Term weighting (*DF)



39

Comparison of best and worst



40

In summary

- The previous experiments allow us to choose the parameters θ^* that most closely accord with the similarity judgements implicitly embodied in ODP
  - term selection:
    - page content
    - size 32 anchor windows
  - weighting schemes:
    - distance
    - NMDF

41

Scaling to large repositories

- Goal: generate a similarity index that allows efficient runtime query processing, using strategy θ^*
- Dataset: 80M URLs from Stanford WebBase

42

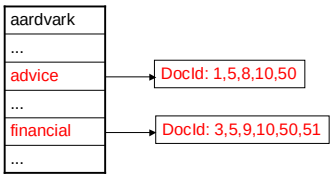
Scalability:
keyword search ≠ similarity search

- For standard keyword search query, # of accesses to inverted index equals # of terms in query
- The postings lists for most terms are of reasonable length

43

Typical keyword search

typical keyword search query: "financial advice"



Inverted index lookup is manageable

44

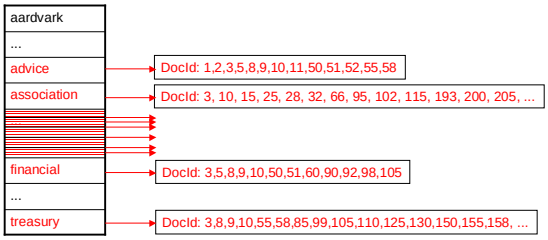
Scalability:
keyword search ≠ similarity search

- For similarity search, # of accesses to inverted index equals # of terms in the query page's (potentially large) bag
- Many of these terms could have huge postings list in the inverted index
  - content words
  - very wide anchor windows

45

Scalability

typical similarity search query: "www.money.com"



Inverted index lookup is ***not*** manageable

46

Scalability

- Solution summary:
  - Use special kind of signature generation technique to represent bags with fixed-length signature vector
  - Similar signature vectors indicate similar bags, w.h.p.
  - [Broder et al STOC '98], [Indyk SODA '99]

47

Sample results

MSN Money
MSN Money
Money Magazine
MoneyExtra
Money
ETrade
Money Club
MorningStar
The Money Page
Reuters MoneyNet
MutualFunds

MP3.com
International Music Network
EMusic
CMJ: New Music First
EMusic
Lycos Music
AudioGalaxy
Listen.com
Launch.com
Nullsoft Winamp
Gracenote (cddb)

48

Sample bags

Top 5 words from each bag are shown

moneycentral.msn.com	money, finance, msn, website, moneycentral
www.weather.com	weather, channel, forecasts, fbc, enter
www.cnnfn.com	finance, business, cnn, cnnfn, stock
www.mp3.com	music, audio, player, artist, napster
java.sun.com	java, jdk, technology, microsystems, api
www.cdnow.com	music, cdnow, amazon, records, books

49

Future work

- What if ODP pages aren't representative of web pages in general?
- Calculate several "best" parameter settings, based on certain page properties
  - Calculate separate Γ scores for strategy over low indegree and high indegree pages
  - Partition scores for other properties as well ...

50