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 \( q \)

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

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

<table>
<thead>
<tr>
<th>Similarity search</th>
<th>Related work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Finding Related Pages in the WWW</td>
</tr>
<tr>
<td></td>
<td>- [Dean,Henzinger WWW8 ’99]</td>
</tr>
<tr>
<td></td>
<td>Automatic Resource Compilation ...</td>
</tr>
<tr>
<td></td>
<td>- [Chakrabarti et al WWW7 ’98]</td>
</tr>
<tr>
<td></td>
<td>Commercial search engines</td>
</tr>
</tbody>
</table>

Similarity search

www.moneycentral.com
www.pathfinder.com/money
www.moneyworld.co.uk
www.money.com
www.etrade.com
www.moneyclub.com
Model for document similarity

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

Model for document similarity

- For pages $a$ and $b$, with respective bags $\alpha$ and $\beta$, define
  \[ \text{sim}(a, b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \]
- Strategy for (page $\rightarrow$ bag) is the crucial step in quality of sim()
Similarity search system

Possible term choices

Content

Links
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, $\theta$, defines a similarity search strategy

Similarity search system

(Strategy, query) $\rightarrow$ similarity ordering

- Inputs:
  - $\theta \in \Theta$: strategy (i.e., parameter setting)
  - $q \in \text{Web}$: query page
- Outputs:
  - $\tau$: list of web pages ordered by similarity to $q$ using strategy $\theta$
  - $\tau = \Gamma(\theta, q)$
Evaluating strategies

- Goal: find “best” \( \theta_1 \in \Theta \)
- 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?

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

Directory \(\rightarrow\) Similarity judgements

Computers

Hardware

www.hardware.com

Software

www.software.com

www.machin.com

www.prog.com

Open Directory

Query

Same Class

Sibling Class

Cousin Class

Unrelated
(Directory, query) → similarity ordering

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

---

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

---

Directory vs. Strategy

- **Open Directory**
  - ODP: strict weak ordering of pages (ODP)
  - τ_i: total ordering of pages according to θ_i

- **Strategy θ_i**
  - Based on Kruskal-Goodman Γ

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 ≤ Γ ≤ 1: measure of agreement
    
    \[ 2 \times \Pr[τ_{odp} \text{ and } τ_i \text{ agree on ordering of } (u,v)] - 1 \]
Example of two rankings with different \( \Gamma \) scores

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

<table>
<thead>
<tr>
<th>Organization</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Botanical Conservation Network</td>
<td><a href="http://www.cbcn.org">http://www.cbcn.org</a></td>
</tr>
<tr>
<td>The Royal Horticultural Society</td>
<td><a href="http://www.rhs.org.uk">www.rhs.org.uk</a></td>
</tr>
<tr>
<td>The American Rhododendron Society</td>
<td><a href="http://www.rhs.org.uk">www.rhs.org.uk</a></td>
</tr>
<tr>
<td>Gardener's Supply Company</td>
<td><a href="http://www.vg.com">www.vg.com</a></td>
</tr>
</tbody>
</table>

\[ \Gamma = 0.5312 \quad \Gamma = 0.3096 \]

Evaluating strategies

1. For each \( \theta_i \in \Theta \)
   \[ \Gamma_{\theta_i} = \text{Avg}_{q \in D} \left[ \Gamma(\Theta, q), \Gamma(\theta_i, q) \right] \]

2. Select strategy \( \theta^* = \arg\max_{\theta_i} [\Gamma_{\theta_i}] \)

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

Feature space: term selection

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

Directory vs. Strategy

Gamma scores
Orthogonality

Fraction of Pairs that are Orthogonal

Directory → Similarity judgements

Computers

Hardware

www.hardware.com

Software

www.software.com

www.machine.com

www.programming.com

Composite schemes

Feature space: term weighting

- Distance weighting for anchor-window terms

Note: distance weighting was enabled
The 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

Term weighting (*DF)

Comparison of best and worst
In summary
- The previous experiments allow us to choose the parameters $\theta^*$ that most closely accord with the similarity judgements implicitly embodied in ODP
  - term selection:
    - page content
    - size 32 anchor windows
  - weighting schemes:
    - distance
    - NMDF

Scaling to large repositories
- Goal: generate a similarity index that allows efficient runtime query processing, using strategy $\theta^*$
- Dataset: 80M URLs from Stanford WebBase

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

Typical keyword search
- typical keyword search query: “financial advice”

<table>
<thead>
<tr>
<th>Term</th>
<th>DocId</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>1, 5, 8, 10, 50</td>
</tr>
<tr>
<td>advice</td>
<td>DocId: 1, 5, 8, 10, 50</td>
</tr>
<tr>
<td>financial</td>
<td>DocId: 3, 5, 9, 10, 50, 51</td>
</tr>
</tbody>
</table>

Inverted index lookup is manageable
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

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]

Sample results

<table>
<thead>
<tr>
<th>MSN Money</th>
<th>MP3.com</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN Money</td>
<td>International Music Network</td>
</tr>
<tr>
<td>Money Magazine</td>
<td>EMusic</td>
</tr>
<tr>
<td>MoneyExtra</td>
<td>CMJ: New Music First</td>
</tr>
<tr>
<td>Money</td>
<td>EMusic</td>
</tr>
<tr>
<td>ETrade</td>
<td>Lycos Music</td>
</tr>
<tr>
<td>Money Club</td>
<td>AudioGalaxy</td>
</tr>
<tr>
<td>MorningStar</td>
<td>Listen.com</td>
</tr>
<tr>
<td>The Money Page</td>
<td>Launch.com</td>
</tr>
<tr>
<td>Reuters MoneyNet</td>
<td>Nulsoft Winamp</td>
</tr>
<tr>
<td>MutualFunds</td>
<td>Gracenote (cddb)</td>
</tr>
</tbody>
</table>
Sample bags

Top 5 words from each bag are shown

<table>
<thead>
<tr>
<th>Sample bag</th>
<th>Top 5 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>moneycentral.msn.com</td>
<td>money, finance, msn, website, moneycentral</td>
</tr>
<tr>
<td><a href="http://www.weather.com">www.weather.com</a></td>
<td>weather, channel, forecasts, fbz, enter</td>
</tr>
<tr>
<td><a href="http://www.cnnfn.com">www.cnnfn.com</a></td>
<td>finance, business, cnn, cnnfn, stock</td>
</tr>
<tr>
<td><a href="http://www.msp3.com">www.msp3.com</a></td>
<td>music, audio, player, artist, napster</td>
</tr>
<tr>
<td>java.sun.com</td>
<td>java, jdk, technology, microsystems, api</td>
</tr>
<tr>
<td><a href="http://www.cdnow.com">www.cdnow.com</a></td>
<td>music, cdnow, amazon, records, books</td>
</tr>
</tbody>
</table>

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 I* scores for strategy over low indegree and high indegree pages
  - Partition scores for other properties as well...