Improved Algorithms for Topic Distillation in a Hyperlinked Environment
(ACM SIGIR ‘98)

Ruey-Lung, Hsiao
Nov 23, 2000

Topic Distillation on the WWW

• Definition
  Given a typical user query to find quality documents related to the query topic.

• Characteristics
  – More general than finding a precise query match
  – Not as ambitious as trying to exactly satisfy user information need
  – In cases where query is ambiguous, it should return relevant documents for (some of) the main query topics.
Related Research

1. Authoritative sources in a hyperlinked environment ‘97
4. Inferring Web Communities from link topology ’98
5. What is this page known for ? Computing Web Page Reputations. ’00

HITS (Hyperlink Induced Topic Search)

• Algorithm
  – Start with a root set S
    • $S_0$ is relatively small (typically up to 200 pages)
    • $S_0$ is rich in relevant pages
    • $S_0$ contains most (or many) of the strongest authorities.
  – Recursively compute the degree of authority and hub for each element.

\[
a(p) = \sum_{q \rightarrow p} h(q)
\]
\[
h(p) = \sum_{p \rightarrow q} a(q)
\]
HITS (Hyperlink Induced Topic Search)

• Premises
  – The implicit annotation provided by human creator contains sufficient information to infer authority.
  – The sufficiently broad topics contain embedded communities of hyperlinked pages.

• Problems
  – Mutually Reinforcing Relationships
certain arrangements of documents “conspire” to dominate the computation.
  – Automatically Generated Links
    no human opinion is expressed by the link.
  – Non-relevant Documents
    the graph contains documents not relevant to the query topic

Improved Algorithm

• Improved Connectivity Analysis
  – Mutually reinforcing relationships should have the same influence on a single document.

\[
a(p) = \sum_{q} h(q) \times \text{auth}_w(q,p)
\]

\[
h(p) = \sum_{q} a(q) \times \text{hub}_w(p,q)
\]

• Pruning Nodes from Neighborhood Graph

\[
\text{Similarity}(Q,D_j) = \sum_{i=1}^{j} \frac{W_{iq} \times W_{pj}}{\left( \sum_{i=1}^{j} W_{iq}^2 \right)^{1/2} \left( \sum_{i=1}^{j} W_{pj}^2 \right)^{1/2}}
\]

  – Relevant threshold :
    • Median Weight
    • Start Set Median Weight
    • Fixed Fraction of Maximum Weight
Partial Content Analysis

- Selectively analyze and prune if needed, the nodes that are most influential in the outcome.
- Query Q formation (use 30 documents)
  
  Heuristic: \( \text{in\_degree} + 2 \times \text{num\_query\_matches} + \text{has\_out\_links} \)

- Pruning
  - Degree Based Pruning
    - Use \( 4 \times \text{in\_degree} + \text{out\_degree} \) as a measure of influence
    - Fetch the top 100 nodes, scored against Q and pruned if needed.
  - Iterative Pruning
    - Use connectivity analysis itself to select nodes to prune.
    - Pruning happens over a sequence of rounds, each runs imp for 10 iterations to get ranked list.

Evaluation

<table>
<thead>
<tr>
<th>Average Precision at Top 5 and 10 ranked authority documents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Without Regulation</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>base</td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td>Rare</td>
</tr>
<tr>
<td>Popular</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Average Precision at Top 5 and 10 ranked hub documents

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Without Regulation</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>base</td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td>Rare</td>
</tr>
<tr>
<td>Popular</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Finding Related Pages in the WWW

• Appears in 8th www conference

• Definition
  – A related web page is one that addresses the same topic as the original page.

• Algorithms
  – Companion algorithm: derived from HITS.
  – Cocitation algorithm: finds pages that are frequently cocited with the input URL u.

• Evaluation
  – Two proposed algorithms are 73% better, 51% better than Netscape’s “What’s Related”.

Companion Algorithm

• Takes as input a URL u and consists of four steps:
  – Build a vicinity graph for u.
  – Contract duplicates and near-duplicates in this graph
  – Compute edge weights based on host to host connections
  – Compute hub/authority score.
Cocitation Algorithm

- **Degree of co-citation**
  - The number of common parents of two nodes.