Usage Aware Average-Clicks

Kalyan Beemanapalli — University of Minnesota
Ramya Rangarajan — University of Minnesota
Jaideep Srivastava — University of Minnesota

Presenter:
Kalyan Beemanapalli
Outline

- Introduction
- Related Work
- Background
- Method
- Experiments and Results
- Key Contributions
- Conclusions and Future Work
- Questions
Related Work – Link Analysis

- Applications
  - PageRank
  - HITS
  - Average-Clicks (*Matsuo et al*)

- Disadvantage
  - Static
Related work

- Solution
  - Usage Data

- Why Usage Aware Average-Clicks?
  - Average-Clicks
    - Fairly new algorithm
    - Proposes a new definition to distance between web pages
    - Measures distance in user’s context
  - Ideas from
    - Usage Aware PageRank (*Oztekin et al*)
    - Extensions to HITS (*Miller et al*)
Average-Clicks

- Measure of distance between web pages
- Definition – An average click is one click among \( n \) links
- Probability of a random surfer on a page \( p \) to click any one of the links is

\[
\frac{\alpha}{\text{OutDegree}(p)},
\]

where \( \alpha = \text{Damping Factor} \)
Average Clicks

- Average Click length of links on page $p = -\log_n(\alpha/\text{OutDegree}(p))$.

Where $\alpha = \text{Damping Factor}$, $n = \text{Average Number of links on a page}$

Distance between page $p$ and $q$

- shortest path between the nodes representing the pages in the graph
- Path through a longer chain of links can be considered shorter than one through smaller number of links
Average Clicks - Example

From A to C, 1.56 average-click, 2 clicks.

From A to D, 0.92 average-click, 2 clicks.
Usage Aware Average-Clicks

Usage Graph

No. of occurrences of each page

No. of co-occurrences of pages

$C(p, q) = \frac{\text{Number of co-occurrences of } p, q}{\text{Number of occurrences of } p}$

$C(p, q) = \frac{\text{Weight of the edge from } p \text{ to } q}{\text{Weight assigned to node } p}$
Usage Aware Average-Clicks

\[ D(i, j) = \frac{1}{\text{Outdegree}(\text{page } i)} \text{ if there is a link to page } j \text{ on page } i \]
\[ \infty \text{ otherwise} \]
Usage Aware Average-Clicks

- We now have

\[ C(p,q) = \frac{\text{Number of co-occurrences of } p,q}{\text{Number of occurrences of } p} \]

\[ D(p,q) = \begin{cases} \frac{1}{\text{Outdegree of } e(\text{page } p)} & \text{if there is a link to page } j \text{ on page } i \\ \infty & \text{otherwise} \end{cases} \]

- We combine the Link Matrix and Usage Matrix to define the new definition of distance between 2 pages as follows:

\[ \text{Distance}(p,q) = (1 - C(p,q))^\alpha \times \log \left( \frac{d}{\text{Outdegree}(p)} \right) \]
Usage Aware Average-Clicks

- Shortest distance between pairs of nodes – all pairs shortest path algorithm
- All Pairs Shortest path algorithm used – Floyd Warshall’s Algorithm
- Implementation Issues
  - Poor scalability
Solution

Template for each node

Set of links for page 0

Data Structure for Floyd Warshall

Template for each node

<table>
<thead>
<tr>
<th>Page ID</th>
<th>Avg Click Score</th>
<th>Usage Score</th>
<th>Usg Avr Avg Click Score</th>
</tr>
</thead>
</table>
Experimental Results

- Experiments conducted on www.cs.umn.edu
- Usage data collected in Apr 2006
- Data set reduced to 100,000 sessions
- Noise removed
- Link Graph built using our crawler
Example Distances

<table>
<thead>
<tr>
<th>Destination Page</th>
<th>Average-Clicks</th>
<th>Usage aware Average-Clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.cs.umn.edu/index.php">http://www.cs.umn.edu/index.php</a></td>
<td>0.0566667</td>
<td>0.000612</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/graduate/evaluation.php">http://www.cs.umn.edu/admissions/graduate/evaluation.php</a></td>
<td>0.0566667</td>
<td>0.002460</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/graduate/procedure.php">http://www.cs.umn.edu/admissions/graduate/procedure.php</a></td>
<td>0.0566667</td>
<td>0.002460</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/graduate/checklist.php">http://www.cs.umn.edu/admissions/graduate/checklist.php</a></td>
<td>0.0566667</td>
<td>0.000612</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/graduate/fellowships.php">http://www.cs.umn.edu/admissions/graduate/fellowships.php</a></td>
<td>0.0566667</td>
<td>0.002460</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/graduate/transfers.php">http://www.cs.umn.edu/admissions/graduate/transfers.php</a></td>
<td>0.0566667</td>
<td>0.056666</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/graduate/application.php">http://www.cs.umn.edu/admissions/graduate/application.php</a></td>
<td>0.0566667</td>
<td>0.003690</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/graduate/faculty.php">http://www.cs.umn.edu/admissions/graduate/faculty.php</a></td>
<td>0.0566667</td>
<td>0.001228</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/about/contact.php">http://www.cs.umn.edu/about/contact.php</a></td>
<td>0.0566667</td>
<td>0.000612</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/admissions/index.php">http://www.cs.umn.edu/admissions/index.php</a></td>
<td>0.0566667</td>
<td>0.000612</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/degrees/grad/index.php">http://www.cs.umn.edu/degrees/grad/index.php</a></td>
<td>0.0566667</td>
<td>0.001228</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/degrees/grad/pha.php">http://www.cs.umn.edu/degrees/grad/pha.php</a></td>
<td>0.0566667</td>
<td>0.056666</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/degrees/grad/ms.php">http://www.cs.umn.edu/degrees/grad/ms.php</a></td>
<td>0.0566667</td>
<td>0.001845</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/degrees/grad/mcs.php">http://www.cs.umn.edu/degrees/grad/mcs.php</a></td>
<td>0.0566667</td>
<td>0.003073</td>
</tr>
<tr>
<td><a href="http://www.cs.umn.edu/research/research_areas.php">http://www.cs.umn.edu/research/research_areas.php</a></td>
<td>0.0566667</td>
<td>0.001228</td>
</tr>
</tbody>
</table>

Comparison of results from Average-Clicks and Usage Aware Average-Clicks
Evaluation Methodology

■ Domain Expert’s View
  • Questionnaires

■ User’s View
  • Questionnaires
  • Automate verification

■ Our Method
  • Predicting Power
Evaluation Methodology

- Incorporated into a recommender system
- Idea - pages that are close to each other are more similar to each other than pages that are farther apart
- Performance compared with ‘2, -1’ model
- Tested on www.cs.umn.edu
The Recommender System Architecture

Web Logs → Session Identification → Sessions → Session Alignment → Session Similarity → Graph Partitioning → Session Clusters

Website → Usage Aware Average-Clicks Generation

Usage Aware Average-Clicks Hierarchy → Clickstream Trees

Get Recommendations → Web Server → Web Client

Webpage request → HTML + Recommendations

Recommendations → Recommendation System

Offline

Online
Evaluation Measures

- **Hit Ratio (HR):** Percentage of hits. If a recommended page is actually requested later in the session, we declare a hit.

- **Click Reduction (CR):** For a test session \((p1, p2, \ldots, pi\ldots, pj\ldots, pn)\), if \(pj\) is recommended at page \(pi\), and \(pj\) is subsequently accessed in the session, then the click reduction due to this recommendation is,

\[
Click\ reduction = \frac{j-i}{i}
\]
Experimental Set-up

- 1000 training sessions
- 3, 5, 10 recommendations
- 10, 15 and 20 ClickStream Clusters
- Different testing sessions
- Experiment repeated 5 times using different training set
- Results compared against the ‘2, -1’ model
- T-tests performed
- Same procedure for 3000 training sessions
Results

Hit Ratio Vs No. of Recommendations for 1000 sessions, 10 clusters

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>p value</td>
<td>0.123242</td>
<td>0.030262</td>
<td>0.006292</td>
</tr>
</tbody>
</table>

t-test scores for 1000 sessions, 10 clusters
Hit Ratio Vs No. of Recommendations for 1000 sessions, 15 clusters

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>p value</td>
<td>0.053543</td>
<td>0.014464</td>
<td>0.020082</td>
</tr>
</tbody>
</table>

t-test scores for 1000 sessions, 15 clusters

Hit Ratio Vs No. of Recommendations for 1000 sessions, 20 clusters

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>3</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>p value</td>
<td>0.04985</td>
<td>0.224891</td>
<td>0.125186</td>
</tr>
</tbody>
</table>

t-test scores for 1000 sessions, 20 clusters
% Path Reduction

% Path Reduction Vs No. of Recommendations for 1000 sessions, 10 clusters
% Path Reduction Vs No. of Recommendations for 1000 sessions, 15 clusters

% Path Reduction Vs No. of Recommendations for 1000 sessions, 20 clusters
Conclusion

- Incorporated usage data into Average Clicks algorithm.
- Proposed a distance model using usage data and link graph
- Used this method to calculate the similarity between the pages in an intranet domain
- Showed that using a combination of web graph and link graph will provide better recommendations
Future Work

- Validate the algorithm using various testing methods like
  - Domain expert testing
  - User’s perspective

- Compare the algorithm against other usage based link analysis algorithms

- Compare the quality of recommendations with those obtained by using other kinds of domain information
Questions