Studying Trailfinding Algorithms for Enhanced Web Search

Adish Singla, Microsoft Bing
Ryen W. White, Microsoft Research
Jeff Huang, University of Washington
IR Focused on Document Retrieval

- Search engines usually return lists of documents
- Documents may be sufficient for known-item tasks
- Documents may only be starting points for exploration in complex tasks
  - See research on orienteering, berrypicking, etc.
Beyond Document Retrieval

- Log data lets us study the search activity of many users
  - Harness wisdom of crowds
  - Search engines already use result clicks extensively

- Toolbar logs also provide non-search engine activity
  - **Trails** from these logs might help future users
  - Trails comprise queries and post-query navigation

- IR systems can return documents *and/or* trails
  - The “trailfinding” challenge
Trailfinding

- Trails can provide guidance to users beyond the results
- Trails can be shown on search result page, e.g.,

**Mount Rainier National Park (U.S. National Park Service)**
Government page about this volcano, environmental information and statistics.

**Mount Rainier - Wikipedia, the free encyclopedia**
Mount Rainier is a large active stratovolcano (also known as a composite volcano) in Pierce County, Washington, USA, located 54 miles (87 km) southeast of Seattle.

- How to select best trail(s) for each query-result pair?
- We present a log-based method and investigation
Outline for Remainder of Talk

- Related work
- Trails
  - Mining Trails
  - Finding Trails
- Study
  - Methods
  - Metrics
  - Findings
- Implications
Related Work

- **Trails as evidence for search engine ranking**
  - *e.g.*, Agichtein et al., 2006; White & Bilenko, 2008; ...

- **Step-by-step guidance for Web navigation**
  - *e.g.*, Joachims et al, 1997; Olston & Chi, 2003; Pandit & Olston, 2007

- **Guided tours (mainly in hypertext community)**
  - Tours are first-class objects, found and presented
  - Human-generated
    - *e.g.*, Trigg, 1988; Zellweger, 1989
  - Automatically-generated
    - *e.g.*, Guinan & Smeaton, 1993; Wheeldon & Levene, 2003
Trail Mining

- Trails sourced from nine months of MSN toolbar logs
- Search trails are initiated by search queries
  - Terminate after 10 actions or 30 minutes of inactivity
- Trails can be represented as Web behavior graphs

Graph properties used for trail finding
Trailfinding Algorithms

- Trailfinding task is defined as:

  Given a query $q$ and an observed click on a trail origin $r$, find the trail $t$ in $T$ with the largest $Score(t, q, r)$

- We can define $Score(t, q, r)$ in a number of ways ...
\[
\text{Score}(t, q, r) = \\
\begin{align*}
\text{Length} & = \text{Number of nodes after origin } r \\
\text{Breadth} & = \text{Number of branches after } r \\
\text{Depth} & = \text{Maximum number of nodes on a single branch from origin } r \\
\text{Frequency} & = \text{Frequency of occurrence of trail } t \text{ for query } q \text{ and origin } r \\
\text{Relevance} & = \left( \sum_{u_x \text{ in } t} \max (\% \text{ query terms in title}_x, \% \text{ query terms in URL}_x) \right) / \text{Length}(t)
\end{align*}
\]
\[ \text{Score}(t, q, r) = \]

- **Trail Diversity**
  - Number of pages in \( t \) with different domain than origin \( r \)

- **Trail Strength**
  - Function of engaging potential of behavior graph and the ease of navigation between trail nodes

**Step 1:** Count overall frequency of each transition in \( t \) (over all trails)

\[
( q, r, < u_x \rightarrow u_y > ) = \sum_{u_x \rightarrow u_y \text{ in } t} \text{Freq}(t, q, r)
\]

**Step 2:** Score \( t \) based on sum of transition frequencies

\[
= \sum_{u_x \rightarrow u_y \text{ in } t} (u_x \rightarrow u_y, q, r)
\]
Study: Research Qs

- **RQ1**: Of the trails and origins, which source: (i) provides more relevant information? (ii) provides more coverage and diversity of the query topic? (iii) provides more useful information?

- **RQ2**: Among trailfinding algorithms: (i) how does the value of best-trails chosen differ? (ii) what is the impact of origin relevance on best-trail value and selection? (iii) what are the effects of query characteristics on best-trail value and selection?

- **RQ3**: In associating trails to unseen queries: (i) how does the value of trails found through query-term matching compare to trails with exact query matches found in logs? (ii) how robust is term matching for longer queries (which may be noisy)?
Study: Research Qs

- **RQ1**: Of the trails and origins, which source: (i) provides more relevant information? (ii) provides more coverage and diversity of the query topic? (iii) provides more useful information?

- **RQ2**: Among trailfinding algorithms: (i) how does the value of best-trails chosen differ? (ii) what is the impact of origin relevance on best-trail value and selection? (iii) what are the effects of query characteristics on best-trail value and selection?

- **RQ3**: In associating trails to unseen queries: (i) how does the value of trails found through query-term matching compare to trails with exact query matches found in logs? (ii) how robust is term matching for longer queries (which may be noisy)?
Study: Data Preparation

- Large random sample of queries from Bing logs
- Queries normalized, etc.
- Labeled trail pages based on Open Directory Project
  - Classification is automatic, based on URL with back-off
  - Coverage of pages is 65%, partial trail labeling is allowed
- Interest models were constructed for queries & trails
  - E.g., for query [triathlon training]:

<table>
<thead>
<tr>
<th>Label</th>
<th>Norm. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top/Sports/Multi_Sports/Triathlon/Training</td>
<td>0.58</td>
</tr>
<tr>
<td>Top/Sports/Multi_Sports/Triathlon/Events</td>
<td>0.21</td>
</tr>
<tr>
<td>Top/Shopping/Sports/Triathlon</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Study: Metrics

- **Coverage**
  - Query interest model built from top Goo/Yah/Bing results
  - Fraction of query interest model covered by trail

- **Diversity**
  - Fraction of unique query interest model labels in trail

- **Relevance**
  - Query-URL relevance scores from human judges (6pt scale)
  - Average relevance score of trail page(s)

- **Utility**
  - One if a trail page has dwell time of 30 seconds or more
    - Fox et al. (2005) showed dwell \( \geq 30 \) secs. indicative of utility
Study: Method

- For each query-result pair:
  - Select the best trail using each trailfinding algorithm
  - Compute each of the metrics

- Split findings by origin relevance
  - Best – origin results with high relevance ratings
  - Worst – origin results with low relevance ratings

- Micro-averaged within each query and macro-averaged across all queries
  - Obtain a single value for each source-metric pair
Findings: Coverage/Diversity

All differences between algorithms were statistically significant (p < .01)
Findings: Coverage/Diversity

All differences between algorithms were statistically significant (p < .01)

Frequent trails are short and may not cover much of query
Findings: Coverage/Diversity

All differences between algorithms were statistically significant (p < .01)

Relevant trails may only cover one aspect of the query topic
Findings: Avg. Relevance Scores

- Decreases rather than increases
- Relevance defined in relation to original query
  - Needs may evolve during trail following

Needs may change most during long trails
Findings: Vary Origin Relevance

- Divided trail data into two buckets:
  - Best origins: trails with highest origin relevance
  - Worst origins: trails with lowest origin relevance

- Trails help most when initial search results are poor
- Trails may not be appropriate for all search results
Implications

- Approach has provided insight into what trailfinding algorithms perform best and when
- Next step: Compare trail presentation methods
- Trails can be presented as:
  - Alternative to result lists
  - Popups shown on hover over results
  - In each caption in addition to the snippet and URL
  - Shown on toolbar as user is browsing
- More work also needed on when to present trails
  - Which queries? Which results? Which query-result pairs?
Summary

-Presented a study of trailfinding algorithms
-Compared relevance, coverage, diversity, utility of trails selected by the algorithms
-Showed:
  - Best-trails outperform average across all trails
  - Differences attributable to algorithm and origin relevance
-Follow-up user studies and large-scale flights planned

See paper for other findings related to effect of query length, trails vs. origins, term-based variants