Assessing the Scenic Route: Measuring the Value of Search Trails in Web Logs

Ryen W. White, Microsoft Research
Jeff Huang, University of Washington
Overview

- IR focuses on effectively ranking **documents**
- Individual pages may be inadequate for complex needs

- Trails are a series of pages starting with a query and terminating with session inactivity

- Little is known about the value that trails bring to users

- **Estimate the benefit of trail following behavior**
  - *Know if/when useful ➔ Build trail-centric IR systems*
Trails and Sources

- **Full-trail** =

- **Origin** = First page in trail (visited from result click)
- **Destination** = Last page in trail
- **Sub-trail** = All trail pages other than destination

*We compare the value of these sources for Web search*
Outline for Remainder of Talk

- Related Work
- Study
  - Research Questions
  - Trail Mining and Labeling
  - Metrics
  - Method
  - Findings
- Implications
- Summary
Related Work

- Trails
  - Trailblazing users create links between docs (Bush, 1945)
- Destinations
  - Teleportation (Teevan et al., 2004; White et al., 2007)
- Information-seeking strategies
  - Orienteering, berrypicking, information foraging, etc.
- Guided tours (hypertext community)
  - Human generated (Trigg, 1988)
  - Automatically generated (Guinan & Smeaton, 1993)
Research Questions

- Which sources (origin, destination, sub-trail, full-trail):
  - Provide more relevant information? \((RQ_1)\)
  - Provide more topic coverage? \((RQ_2)\)
  - Provide more topic diversity? \((RQ_3)\)
  - Provide more novel information? \((RQ_4)\)
  - Provide more useful information? \((RQ_5)\)

- Answers help us understand the value of trail following compared with viewing only origin and/or destination
Millions of trails were mined from MSN toolbar logs
  - March – May 2009, 100K unique users
  - Trails comprise queries and post-query behavior

Labeled trail pages based on Open Directory Project
  - Classification is automatic, based on URL with backoff
  - Coverage of pages is 65%, partial trail labeling is allowed

Normalized queries, trails $\geq 3$ URLs, 10 trails/user max, ...
Metrics

- Relevance, Utility, Coverage, Diversity, Novelty
- Per-source (origin, destination, sub-trail, full-trail)
- Trails from May 2009 used for evaluation
  - Trails from March/April used for novelty metric (more later)

- Metric values from sub-/full-trails may be larger than origins/destinations
  - Sub-trails / full-trails contain more pages
  - Extent of the difference important if we are going to show trails on the SERP (large benefit => support to include)
Metrics: Relevance/Utility

- **Relevance**
  - 6-pt Query-URL relevance scores from human judges
  - Average relevance score of sources

- **Utility**
  - One if source has dwell time of 30 seconds or more
    - Supporting evidence from Fox et al. (2005)

- ... also Coverage, Diversity, and Novelty
Metrics: Topic Coverage

- First need ground truth
  - Query interest model built from union of top GYB results

<table>
<thead>
<tr>
<th>Label</th>
<th>( w_l )</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>

- Fraction of query interest model \((q_x)\) covered by source
- For each trail source \((t_s)\), build interest model, compute:

\[
Coverage(t_s) = \sum_{l \in (s \cap q_x)} w_l
\]
Metrics: Topic Diversity

- Fraction unique query-relevant concepts from source
- Same procedure as coverage
  - Build source interest model, compare w/ query model ($q_x$)
  - Fraction unique labels in $q_x$, not fraction of $q_x$

For each trail source $t_s$:

$$Diversity(t_s) = \sum_{l \in (s \cap q_x)} \frac{1}{|q_x|}$$
Metrics: Novelty

- Amount of new query-relevant info. encountered
  - Depends on both query and user
  - What’s novel to one user, may not be to another
- Build historic interest model \((h_x)\) from Mar-Apr logs
  - Summarizes previous query-relevant interests for each user
- Count unique labels in source, \(q_x\), and not in \(h_x\)
- For each source \(t_s\):

\[
Novelty(t_s) = \sum_{l \in (s \cap q_x) \land l \notin h_x} \frac{1}{|q_x|}
\]
Methodology

- Construct query interest models $Q_x$ based on set of 8K queries with human relevance judgments
- Construct historic interest models for each user-query pair in Mar-Apr, filtered to only include queries appearing in $Q_x$
- For each search trail $t$ in $T_x$:
  - Assign ODP labels to pages all pages in $t$
  - Build source interest models for all sources
- Compute relevance, coverage, diversity, novelty, utility
- Average per query, and then average across all queries
Findings: All Queries

Normalized relevance scores range from 0-1 for talk, not study.

Note: Normalized relevance scores range from 0-1 for talk, not study.

All differences significant at $p < .01$
Findings: All Queries

Relevance is lower for trails

Users may be wandering to non-query relevant pages during the trail

Return to relevant destinations at the end of the trail

Note: Normalized relevance scores to range from 0-1 for talk, not study
Findings: Query Popularity

- Grouped queries into Low, Med, High popularity
- Query popularity increase, metrics increase (all sources)
- For example, for normalized relevance:

  ![Normalized Relevance Chart]

- Re: relevance - engine performance improves w/ query popularity (see Downey et al., 2008)
Findings: Query History

- Effect of user “re-finding” on source value
- Grouped queries into:
  - *None*: No instances for the user
  - *Some*: On average ≤ 30 instances in user history
  - *Lots*: On average > 30 instances in user history

- Relevance, utility rise – users more familiar with topic
- Coverage, diversity, novelty fall
  - Less variance in pages visited (see Tyler & Teevan, 2010)
Implications

• Nature of query is important in source value
  • Popularity, history, ... should be considered
  • Trails might be useful for supporting exploration
  • Trails might be a hindrance for focused tasks

• How to select trails
  • Maximizing metrics presented in this talk?
  • Personalizing to user’s search history?
  • Recommend trails when destination unclear?

• How and when to integrate trails into search results
• Follow-up user studies and large-scale flights needed
Summary

- Presented study estimating the value of trails to users
- Systematically compare value of trails to other sources
- Evaluation showed that trails provide significantly more coverage, diversity, novelty, etc. than pages
  - Findings vary by query popularity / re-finding

Next steps:
- Investigate best-trail selection
  - See other SIGIR 2010 paper (Singla, White, Huang)
- Incorporate trails into search engine result pages