Evaluating the ‘Finding’ Experience: 

*Test the Process, not the Result*

Raman Chandrasekar  
Researcher, Microsoft Corporation  
[RamanC@microsoft.com](mailto:RamanC@microsoft.com)

*WWW 2002, Hawaiʻi*
Web Experiments & Test Collections: Are they meaningful?

- Sort of, in a limited fashion.

- Can be made more meaningful with a little effort!

- Focus here: relevance testing.
What is Relevance, anyway?

- Hard question. One of those “I cannot define it, but I recognize it when I see it” issues?
- Need a handle on relevance, to provide a great finding experience
- Kinds of Relevance
  - Textual Relevance
  - Conceptual Relevance
  - Utility ...
    - Examples: gateway, Microsoft, Java
Relevance Testing Schemes

- Ad-hoc one-query ‘tests’
- Pseudo-scientific 5 query ‘tests’
- CNET’s ‘Search Site Olympics’
- eTesting Labs tests
- TREC tests (Web Track)
- Search Engine internal relevance tests
Current Relevance Testing

- General assumption:
  Search: query $\rightarrow \{URL\}$
- Search treated as a single step process
- Relevance measured as a function of the result: the presence and position of ‘expected’ URLs in the result set
What’s wrong with this? ...1

- Ignores HCI research that shows information finding is an iterative process, even for known-item searching
  - So it’s not much use checking results at the first instance.
- Ignores richness, presentation of result page
- Ignores human ability to skip over irrelevant information, and zoom to relevant information
- Ignores difficulties in creating a gold standard “Expected URLs” list
  - intents vary, redirects confuse, the web is dynamic ...

... and more...
What’s wrong with this? ...2

- No consistent definition of a ‘result’
  - Is a relevant ad a result? Sponsored sites? News?

- No way to give credit for features that help in information-finding:
  - popular search topics, spelling correction, cached pages, clustered folders, category links...

- No way to reward/‘punish’ for UI
So: What should we do?

- The central problem in web search: *Satisfying users’ web information finding needs*
- The test: *Are we satisfying the user?*

We propose:

*Process-based evaluation of ‘finding’*
Process-based Evaluation

Informal definition:

- Follow user behavior from query till the user finds a satisfactory result, or until she gives up.
- Compute a satisfaction score based on the ‘cost’ of getting to the result.
How? Queries & judgments

- Blend random queries obtained from several search engines
- Get a bunch of users to ‘find’ information for each query they’re familiar with.
- Track user’s interactions, recording every click → ‘user sessions’. [privacy concerns]
  - Not difficult, we have a prototype for this. Nothing special required for any ‘engine’.
  - Or use something like the Google toolbar

Note: Intent may vary across the process.
Example: India (Arie)

Query → Disambiguation → Site → Satisfaction!
Example: Aaliyah

Query → Auto Spell Correct → Site → Satisfaction!
Sample session data

<table>
<thead>
<tr>
<th>Id</th>
<th>User</th>
<th>Query</th>
<th>Date/Time</th>
<th>URL/code</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Chandra</td>
<td>India</td>
<td>Mon May 6 15:03:31 2002</td>
<td>STARTUP</td>
</tr>
<tr>
<td>11</td>
<td>Chandra</td>
<td>India</td>
<td>Mon May 6 15:03:41 2002</td>
<td><a href="http://search.msn.com/results.asp?co=15.20&amp;ba=0&amp;cfg=SMCINITIAL&amp;v=1&amp;FORM=EQRA&amp;q=India">http://search.msn.com/results.asp?co=15.20&amp;ba=0&amp;cfg=SMCINITIAL&amp;v=1&amp;FORM=EQRA&amp;q=India</a></td>
</tr>
<tr>
<td>11</td>
<td>Chandra</td>
<td>India</td>
<td>Mon May 6 15:03:50 2002</td>
<td><a href="http://www.indiaarie.com">http://www.indiaarie.com</a></td>
</tr>
<tr>
<td>11</td>
<td>Chandra</td>
<td>India</td>
<td>Mon May 6 15:03:55 2002</td>
<td>DONE</td>
</tr>
<tr>
<td>12</td>
<td>Chandra</td>
<td>aliyah</td>
<td>Mon May 6 15:04:10 2002</td>
<td>STARTUP</td>
</tr>
<tr>
<td>12</td>
<td>Chandra</td>
<td>aliyah</td>
<td>Mon May 6 15:04:23 2002</td>
<td>DONE</td>
</tr>
</tbody>
</table>
How? Query Session Analysis

- Define a cost for each step: spelling correction, query modification, give-ups ...
  - e.g. autospell is good, so: a negative cost
- Compute a cost for the query as a whole.
- Compute a satisfaction score for an engine from query costs, averaged over several queries and users
- Relevance proportional to satisfaction score.
Do we need a testing corpus?

- Depends.
  - Scalability and performance critical in Web Search; not replicable in small(er) test collections, makes testing less meaningful.
  - No special testing corpus required for process-based evaluation of ‘finding’.
  - However, a test corpus can help distinguish between technology and content contributions, but ...
Testing Collection: Some Issues

- **Size**: What’s a big enough corpus that’s small enough to share?

- **Type**: Random nodes or a reasonably connected sub-graph? Recent or old? One language or many?

- **Representativeness**: must account for spam, connectivity, weirdnesses.
Summary

- Current relevance testing is limited in many ways.
- Process-based evaluation of ‘finding’ can obviate many current problems.
Mahalo, Aloha!

- These ideas have grown out of discussions within MSN Search.

- Special thanks to:
  - Tom White, Susan Dumais, Philip Carmichael, Susan Dziadosz, David Billick, Matthew Dubeck, Ray Sun, Bill Bliss & John Krass
  - All our users!

http://search.msn.com