Motivation

1. Log based evaluation
   - Relevance determined implicitly or by clicks
   - Large scale collection of user data
   - Suffers from several (click, presentation etc) biases.

2. Batch evaluation
   - Relevance labels assigned by trained judges
   - Smaller test collections
   - Simple assumptions about real information needs

These two forms of evaluation often do not completely agree with each other ([1] and [2]). They agree with each other only when there is a significant gap in quality of the systems compared ([3] and [4]).

User Model

1. When users first access the page, they quickly scan it to determine portions relevant to the query.
   **FINDABILITY**

2. This is followed by reading these paragraphs/snippets.
   **READABILITY**

3. Finally, users focus on understanding these nuggets of information.
   **UNDERSTANDABILITY**

Contributions

1. Identify factors that characterize user effort.
2. Conduct experiments to obtain explicit judgments for these factors.
3. Finally, analyze the effect of incorporating effort into retrieval evaluation.

Methodology

- Collect effort based (explicit) judgments for each document for above parameters.
- Study user preferences
- **Control for relevance**: Collect user preferences with side-by-side comparison for documents of same relevance grade.

Judging Interfaces

- **Trained Judges**
- **End Users**

- At present, relevance is primary factor for judging documents. It does not consider "User effort" (Yilmaz et al. [5])

- A judge can spend a lot of time evaluating correctness of document for a given query. An impatient user may not spend as much time studying the document!

Results

**Factors important for User Satisfaction**

<table>
<thead>
<tr>
<th>FACTOR</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findability*</td>
<td>0.003</td>
</tr>
<tr>
<td>Readability</td>
<td>0.364</td>
</tr>
<tr>
<td>Understandability*</td>
<td>0.054</td>
</tr>
<tr>
<td>Relevance</td>
<td>0</td>
</tr>
</tbody>
</table>

Effort and Preference Agreement

<table>
<thead>
<tr>
<th>FACTOR</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findability</td>
<td>0.60*</td>
</tr>
<tr>
<td>Readability</td>
<td>0.51</td>
</tr>
<tr>
<td>Understandability</td>
<td>0.51</td>
</tr>
<tr>
<td>Relevance</td>
<td>0.72*</td>
</tr>
</tbody>
</table>

Features

**Test Features**

- `avgSumChar` Avg #chars in summary
- `docCLI` CLI Index of document
- `docWords` #words in document
- `qTermsInTitle` #query terms in title
- `sumWords` #words in summary
- `tRatio` Fraction of #words and #tags in html

**Structure Oriented Features**

- `fTable` Fraction of Tables
- `maxWinPos` Max window pos with all query terms
- `qWinO` Fraction of outlinks with query terms
- `fBoldItalics` Fraction of bold, italics and strong
- `sumWords` Fraction of images
- `meanPosOut` Median window pos with all query terms
- `countH` #Headings with query terms

Findability Prediction

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>tTable</code></td>
<td>0.00</td>
</tr>
<tr>
<td><code>avgSumChar</code></td>
<td>0.01</td>
</tr>
<tr>
<td><code>docCLI</code></td>
<td>0.02</td>
</tr>
<tr>
<td><code>maxWinPos</code></td>
<td>0.04</td>
</tr>
</tbody>
</table>

Relevance Prediction

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>tRatio</code></td>
<td>0.01</td>
</tr>
<tr>
<td><code>qWinO</code></td>
<td>0.01</td>
</tr>
<tr>
<td><code>maxWinPos</code></td>
<td>0.04</td>
</tr>
<tr>
<td><code>docWords</code></td>
<td>0.04</td>
</tr>
</tbody>
</table>

Relevant vs Low effort Relevant Documents (p@10)

References

3. J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? In Proc. SIGIR, Salvador, Brazil, 2006.